Technical Report for Cosine

ACM Reference Format:
. 2019. Technical Report for Cosine. In *Proceedings of* . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 DEFINITION OF DISTRIBUTION-DEPENDENT PARAMETERS

We now describe the distribution-aware model in more detail. In this section, we focus on stating the expressions used in the model, and in the next section, we prove the desired expressions capture the quantities that we want.

We use the following notation to describe the number of keys living in run in a hot level i or the total number of keys residing as data in the nodes of a cold level i:

$$R^0 = E_B,\ R^i = \frac{E_B \cdot T^i}{K} \text{ if } 1 \le i \le L-Y-1$$

$$R^L = \frac{E_B \cdot T^i}{Z},\ R^i = \frac{E_B \cdot T^i}{Z} \cdot \frac{B-T}{B} \text{ else if } L-Y \le i \le L-1.$$

We begin by defining the necessary parameters for the lookup cost. For uniform distribution, we define C_0 and $C_{r,i}$. The building blocks of the parameters are the α^i values, which captures the probability that a key k appears in a run (or as data in a node) at level i, and α^{BC}, which captures the probability k is in the block cache. We shows in the next section that $\alpha^i = \min(1, \frac{R^i}{|U|})$ and $\alpha^{BC} = \min(1, \frac{E_{BC}}{|U|})$. We define the next four probabilities $q, c, c_{r,i}, d_{r,i}$ in terms of these building blocks. These building blocks will be explained in the next section.

$$q = 1 - (1-\alpha^0)\left(\prod_{i=1}^{L-Y-1}(1-\alpha^i)^K\right)\left(\prod_{i=L-Y}^{L}(1-\alpha^i)^Z\right)$$

$$c = (1-\alpha^{BC})(1-\alpha^0)\left[1 - \left(\prod_{i=1}^{L-Y-1}(1-\alpha^i)^K\right)\left(\prod_{i=L-Y}^{L}(1-\alpha^i)^Z\right)\right]$$

$$c_{r,i} = (1-\alpha^0)(1-\alpha^{BC})\left(\prod_{h=1}^{i-1}(1-\alpha^h)^K\right)(1-\alpha^i)^r \left(1-(1-\alpha^i)^{K-r}\left(\prod_{h=i+1}^{L-Y-1}(1-\alpha^h)^K\right)\left(\prod_{h=L-Y}^{L}(1-\alpha^h)^Z\right)\right)$$

$$d_{r,i} = (1-\alpha^{BC})(1-\alpha^0)\left(\prod_{h=1}^{L-Y-1}(1-\alpha^h)^K\right)\left(\prod_{h=L-Y}^{i-1}(1-\alpha^h)^Z\right) (1-\alpha^i)^r\left(1-(1-\alpha^i)^{Z-r}\left(\prod_{h=i+1}^{L}(1-\alpha^h)^Z\right)\right).$$

These building blocks fit together to form $C_0 = c/q$, $C_{r,i} = c_{r,i}/q$ for $1 \le i \le L-Y-1$, and $C_{r,i} = d_{r,i}/q$ for $i > L-Y-1$.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

, ,
© 2019 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Next, we define the analogous distribution-dependent parameters C_0^l and $C_{r,i}^l$ for the skew distribution, where $l = 1$ represents a special key and $l = 2$ represents a regular key. The building blocks are similarly $\alpha^{i,1} = 1 - \left(1 - \frac{p_{put}}{|U_1|}\right)^{R^i}$, $\alpha^{i,2} = \min(1, \frac{R^i(1-p_{put})}{|U_2|})$, $\alpha^{BC,1} = 1 - \left(1 - \frac{p_{get}}{|U_1|}\right)^{E_{BC}/B}$, and $\alpha^{BC,2} = \min(1, \frac{E_{BC}(1-p_{get})}{|U_2|})$. We analogously define the four probabilities $q^l, c^l, c_{r,i}^l, d_{r,i}^l$ in terms of these building blocks:

$$q^l = 1 - (1-\alpha^{0,l})\left(\prod_{i=1}^{L-Y-1}(1-\alpha^{i,l})^K\right)\left(\prod_{i=L-Y}^{L}(1-\alpha^{i,l})^Z\right)$$

$$c^l = (1-\alpha^{BC,l})(1-\alpha^{0,l}) \left[1 - \left(\prod_{i=1}^{L-Y-1}(1-\alpha^{i,l})^K\right)\left(\prod_{i=L-Y}^{L}(1-\alpha^{i}(l))^Z\right)\right]$$

$$c_{r,i}^l = (1-\alpha^{0,l})(1-\alpha^{BC,l})\left(\prod_{h=1}^{i-1}(1-\alpha^{h,l})^K\right)(1-\alpha^{i,l})^r \left(1-(1-\alpha^{i,l})^{K-r}\left(\prod_{h=i+1}^{L-Y-1}(1-\alpha^{h}(l))^K\right)\left(\prod_{h=L-Y}^{L}(1-\alpha^{h,l})^Z\right)\right)$$

$$d_{r,i}^l = (1-\alpha^{0,l})\left(\prod_{h=1}^{L-Y-1}(1-\alpha^{h}(l))^K\right)\left(\prod_{h=L-Y}^{i-1}(1-\alpha^{h,l})^Z\right) (1-\alpha^{BC,l}(1-\alpha^{i,l})^r\left(1-(1-\alpha^{i,l})^{Z-r}\left(\prod_{h=i+1}^{L}(1-\alpha^{h,l})^Z\right)\right)$$

These building blocks fit together to form $C_0^l = c^l/q^l$, $C_{r,i}^l = c_{r,i}^l/q^l$ for $1 \le i \le L-Y-1$, and $C_{r,i}^l = d_{r,i}^l/q^l$ for $i > L-Y-1$.

Next, we define the distribution-dependent parameters Q_i for the update cost. The building block of these parameters is the function family $f(V, M) = V \cdot \log\left(\frac{V}{V-M+1}\right)$, which we explain in the next section. This is closely related to Q_i, the number of queries needed to fill up a level, as follows. For a uniform distribution, we let Q_0 be $\max(R^0, f(U, R^0))$ and Q_i be $\max(R^i \cdot K, K \cdot f(U, R^i))$ for $i \neq 0$. When the distribution is skew, the bound is a bit more complex. We define $g_1(0) = f(U_2 + U_1, R^0)$ and $g_1(i) = K \cdot f(U_2 + U_1, R^i)$ for $i \neq 0$. We define $g_2(i)$ to be $-\infty$ if $R^i \le U_1$. If $R^i > U_1$, then we define $g_2(0) = f(U_2, R^0 - U_1)/(1 - p_{put})$ and $g_2(i) = K \cdot f(U_2, R^i - U_1)/(1 - p_{put})$. Now, we define Q_0 to be $\max(R^0, g_1(0), g_2(0))$ and for $i \neq 0$, Q_i is defined to be $\max(R^i \cdot K, g_1(i), g_2(i))$ for skew distributions.

2 PROOFS OF THE DERIVATIONS

2.1 Single-result lookups

For single-result lookups, let $\mathscr{C}$ be the random variable for the cost of a single-result lookup of a given key k. Observe the support of $\mathscr{C}$ is in $\mathbb{Z}^{\ge 0}$. This cost can be decomposed into $\mathscr{C}_0 + \mathscr{C}_{other}$. Here, the random variable C_0 is the binary random variable (support in $\{0, 1\}$) for the I/O cost of accessing the actual (key, value) pair (which we know exists since this is a single-result lookup). The random variable $\mathscr{C}_{other}$ is a random variable with support in $\mathbb{N}$ for the number of other blocks that are touched (from cascading fence pointers or bloom filter false positives). By the setup of the

in-memory helpers and data structure, we know that at most one block in each run at each level is touched. Thus, we can decompose C_{other} into $\sum_{i=1}^{L-Y-1}\sum_{r=1}^{K}\mathscr{C}_{r,i} + \sum_{i=L-Y}^{L}\sum_{r=1}^{Z}\mathscr{C}_{r,i}$, where $\mathscr{C}_{r,i}$ is a binary random variable that keeps whether a block at run r and level i is accessed for anything except for the actual (key, value) pair. Thus, we can write

$$\mathscr{C} = \mathscr{C}_0 + \sum_{i=1}^{L-Y-1}\sum_{r=1}^{K}\mathscr{C}_{r,i} + \sum_{i=L-Y}^{L}\sum_{r=1}^{Z}\mathscr{C}_{r,i}.$$

Using linearity of expectation (which applies to dependent random variables), we know that

$$\mathbb{E}[\mathscr{C}] = \mathbb{E}[\mathscr{C}_0] + \sum_{i=1}^{L-Y-1}\sum_{r=1}^{K}\mathbb{E}[\mathscr{C}_{r,i}] + \sum_{i=L-Y}^{L}\sum_{r=1}^{Z}\mathbb{E}[\mathscr{C}_{r,i}].$$

There are two sources of randomness: (a) the bloom filters, and (b) the data structure layout (i.e. which keys reside in which run). We use the fact that the bloom filters have independent randomness from the data structural layout randomness (which is purely determined the randomness over put queries, $\mathcal{D}_{put}$). Thus, $\mathbb{E}[\mathscr{C}_{r,i}] = p_i \cdot C_{r,i}$ where $C_{r,i} := \mathbb{E}[\mathscr{C}'_{r,i}]$ is the expectation of the binary random variable $C'_{r,i}$ which is 1 if a bloom filter for run r and level i is accessed *or* if a block in run r and level i is accessed, and is 0 otherwise. If no bloom filter exists, then $C_{r,i} := \mathbb{E}[\mathscr{C}_{r,i}]$. Let $C_0 := \mathbb{E}[\mathscr{C}_0]$. We see that this means that:

$$\mathbb{E}[\mathscr{C}] = C_0 + \sum_{i=1}^{L-Y-1} p_i \sum_{r=1}^{K} C_{r,i} + p_i \sum_{r=1}^{Z} C_{r,L-Y} + \sum_{i=L-Y}^{L}\sum_{r=1}^{Z} C_{r,i}.$$

Now, we can simply take an expectation over the randomness in the get distribution to arrive at the expressions in the main paper: when $\mathcal{D}_{get}$ is uniform, we know that $\mathbb{E}[\mathscr{C}]$ is independent of the key k, so we obtain $C_0 + \sum_{i=1}^{L-Y-1} p_i \sum_{r=1}^{K} C_{r,i} + p_i \sum_{r=1}^{Z} C_{r,L-Y} + \sum_{i=L-Y}^{L}\sum_{r=1}^{Z} C_{r,i}$. When $\mathcal{D}_{get}$ is skew, the cost is p_{get} times the expected cost of a special key access plus $1 - p_{get}$ times the cost of a normal key access, as shown in the table.

Now, we need to compute C_0 and each $C_{r,i}$ value for a given key k. Note that $C_0 = \mathbb{E}[\mathscr{C}_0]$, for $i \leq L - Y$, $C_{r,i} = \mathbb{E}[\mathscr{C}'_{r,i}]$, and for $i > L - Y$, $C_{r,i} = \mathbb{E}[\mathscr{C}_{r,i}]$. The tricky part is that by considering a single-result lookup, we are implicitly *conditioning* on the event that key k being in the data structure. So, when evaluating these expectations, we're really considering random variables that live on the "conditional" probability space where the sample space is the event that k is in the data structure. To evaluate the conditional probabilities, we use the fact for a random variable X and an event E, it is true that $\mathbb{P}[X = 1 \mid E] = \frac{\mathbb{P}[X=1 \text{ and } E]}{\mathbb{P}[E]}$. Now, we can formally define the We building blocks $q, c, c_{r,i}, d_{r,i}$ from the previous section. We let q be the probability that key k is in the data structure. We let c be the probability that k is not in memory and k is actually on disk. We use $c_{r,i}$ and $d_{r,i}$ to differentiate between levels with Z runs and levels with K runs. (Note: in this paper, we are making the assumptions that if $Y > 0$, then $Z = 1$). We let $c_{r,i}$ be the probability for $i \leq L - Y - 1$ that the bloom filter at run r at level i is accessed for something other than key k and that key k lives later in the data structure. We let $d_{r,i}$ be the analogous probability for $i > L - Y - 1$. This formulation makes it clear that these building blocks fit together to form $C_0 = c/q$, $C_{r,i} = c_{r,i}/q$ for $1 \leq i \leq L - Y - 1$, and $C_{r,i} = d_{r,i}/q$ for $i > L - Y - 1$.

Thus, it suffices to explicitly compute $q, c, c_{r,i}, d_{r,i}$. At this stage in the computation, everything becomes approximate as the true expressions are intractable to compute exactly. We make several simplifying assumptions throughout the analysis in order to obtain closed-formed estimates of these quantities.

We compute these quantities in terms of fundamental building blocks: the α^i values captures the probability that a key k appears in a run (or as data in a node) at level i, and α^{BC} captures the probability k is in the block cache. For uniform distributions, it is easy to see that $\alpha^i = \min(1, \frac{R^i}{|U|})$ and $\alpha^{BC} = \min(1, \frac{E_{BC}}{|U|})$. For skew distributions, we claim that $\alpha^{i,1} = 1 - \left(1 - \frac{p_{put}}{|U_1|}\right)^{R^i}$ and $\alpha^{BC,1} = 1 - \left(1 - \frac{p_{get}}{|U_1|}\right)^{E_{BC}/B}$ if k is a special key, while $\alpha^{i,2} = \min(1, \frac{R^i(1-p_{put})}{|U_2|})$ and $\alpha^{BC,2} = \min(1, \frac{E_{BC}(1-p_{get})}{|U_2|})$ if k is a normal key. For the special keys, this is actually an appropximation: we ignore that each run / data in a node consists of distinct elements. Now, $1 - \frac{p_{put}}{|U_1|}$ is the probability that a given entry does not contain key k, and we take this to the power R^i to compute an approximate probability that k appears in none of the R^i data entries. (This is slightly inaccurate for cold levels, where the nodes in different parts of the tree contain keys in different intervals, but we ignore this complexity in this work). We then subtract this quantity from 1 to obtain the probability that key k appears. For normal keys, the calculation looks like a lot like the regular uniform calculations with the approximation that only a $(1 - p_{put})$ fraction of the R^i entries will be normal keys (i.e. the rest will be special keys). Note that all of the α^{BC} estimates are slightly inaccurate in that they do not take into account that keys in the block cache are repeated elsewhere in the data structure, but we believe the resulting error is minimal. Moreover, these estimates assume that the data structure is full (i.e. every run is filled to capacity). It is straightforward to add weights w if we instead want to hypothesize that runs are partially filled, and we often use a weight of 1/2 in our predictions.

Now, we show how to aggregate the α values into estimates of q, c, $c_{r,i}$, and $d_{r,i}$. The quantity $(1 - \alpha^{BC})$ gives us an approximate probability that k is not in the block cache, while the quantity $(1 - \alpha^i)$ gives us an approximate probability that k is not in run r / not data in a node at level i. We use the fact that $c_{r,i}$ and $d_{r,i}$ can be computed through considering the probability key k is not at the given run / node or anywhere earlier in the tree and that key k appears later in the tree. This exact expression is intractable so we make several simplifying assumptions. First, we condition on the event that the data structure is full (i.e. every level is filled to capacity), which enables us to use the α estimates to predict the probability that a key appears at a given level. Second, we assume that the merging strategy is such that all of a given level is merged at once, so that the data at runs is (conditionally) "independent", which enables us to treat the probability that key k appears at two different levels as independent. Now, to compute q, we can simply multiply together the probabilities that k is not in each run / node in a level and then subtract this quantity from 1. Similar logic yields c, $c_{r,i}$ and $d_{r,i}$.

2.2 Writes

For update cost, we emphasize the cost estimates are amortized, since any given merge is quite expensive, and merges happen infrequently. We also assume that merges are all-to-one in the sense that all runs in a given run are merged at once into a single run in the next level, and that all of the keys in level $L - Y - 1$ are merged through the cold levels in a batch. For cold levels, we consider how data is propagated from buffers at internal nodes to the final leaf node, but do not consider balancing costs or the costs of the tree shape changing drastically. This is representative of a "small" number of put queries that take place after bulk loading. It suffices to compute the total cost of a key propagating eventually to the bottom level of the data structure, in the following sense: if a put query of key k participates in a merge with keys arising from P other put queries that costs M I/Os, then we say that the cost incurred by a single put query is $\frac{P}{M}$. To compute the cost incurred by a put query, the relevant quantity is $\sum_{i=1}^{L} \mathscr{P}_i$, where $\mathscr{P}$ is the random variable cost incurred by a single put query in the merge from level $i-1$ to level i. For $i \leq L - Y - 1$, this depends on the number of put queries needed to fill level $i-1$. We let this quantity be $\mathscr{Q}_i$. Then it is clear that $\mathscr{P}_i = \frac{R^{i-1} \cdot K + R^i}{B\mathscr{Q}_i}$, since $\frac{R^{i-1} \cdot K + R^i}{B}$ is the read cost. We make the (crude) approximation that $\mathbb{E}[\mathscr{P}_i]$ is approximately equal to $\frac{R^{i-1} \cdot K + R^i}{B\mathbb{E}[\mathscr{Q}_i]}$. For the merges involving cold levels, first let's consider the case of $T = B$. we see that each key in level $L - Y - 1$ results in at most one block access at each future level. Another bound is the total number of blocks in that level. Thus, we obtain $\frac{\min(E_B \cdot T^{L-Y-1}, E_B \cdot T^i/B)}{\mathbb{E}[\mathscr{Q}_i]}$. For $T < B$, there is some gain for $i \geq L - Y + 1$ obtained by buffers batching writes at each internal node. In a given batch of $B - T$ entries, $\min(B - T, T)$ nodes at the next level are touched. So, each write costs at most $\frac{\min(B-T,T)}{B-T}$. This gives a $\frac{\min(B-T,T)}{B-T}$ multiplier on the above expression for levels $i \geq L - Y + 1$.

Thus, it suffices to lower bound $Q_i = \mathbb{E}[\mathscr{Q}_i]$. The building block of these parameters is the function family $f(V, M) = V \cdot \log\left(\frac{V}{V-M+1}\right)$, which is a coupon collector estimate of the number of queries to obtain M distinct elements in a uniform distribution with width V. The coupon collector problem studies the number of distinct values obtained from a sequence of i.i.d draws from a uniform distribution. The closed form estimate $f(V, M)$ is well-known, and is an approximation of the quantity $\frac{V}{V} + \frac{V}{V-1} + \ldots + \frac{V}{V-M+1}$ which captures the exact expected number of queries. This is closely related to Q_i, the number of queries needed to fill up a level, as follows. For a uniform distribution, we can thus compute Q_0 as approximately $f(U, R^0)$ (with some tweaking to account for error in estimation). We can compute Q_i similarly by approximately $K \cdot f(U, R^i)$ for $i \neq 0$. When the distribution is skew, the bound is a bit more complex. One bound arises from the fact we can upper bound by the case of uniform distributions with universe size $U_1 + U_2$ to obtain $g_1(0) = f(U_2 + U_1, R^0)$ and $g_1(i) = K \cdot f(U_2 + U_1, R^i)$ for $i \neq 0$. The other bound arises from the fact that if $R^i > U_1$, then $R^i - U_1$ are necessarily filled up by normal keys, which only show up in a fraction of put queries: we obtain $g_2(0) = f(U_2, R^0 - U_1)/(1 - p_{put})$ and $g_2(i) = K \cdot f(U_2, R^i - U_1)/(1 - p_{put})$. Combining these two estimates with a trivial estimate of M yields the desired expressions.

2.3 No-result lookups

Early stopping does not occur for no-resulting lookups, so every bloom filter must be touched. The number of I/Os now only depends on the randomness of bloom filters.

2.4 Range queries

Every run in a hot level is necessarily touched. The number of blocks touched depends on the selectivity s, and it becomes critical that the runs are sorted. For the cold levels, if the data structure has $T = B$, then we can follow the path down to the leaf nodes and then use the fact that the leaf nodes are connected in a linked list to obtain the desired estimate. If $T < B$, then we need to check internal nodes as well as the leaf nodes for data. The fraction of internal nodes that are touched also depends on the selectivity.

3 ANALYSIS OF DISTRIBUTION-AWARE COST MODELS

We illustrate a comparative analysis of the proposed distribution-aware cost model with the worst-case model using an example. We present a case study of a database with 10 billion (10^{10}) entries experiencing a workload comprising of 50 billion operations. The workload is read-intensive with 80% reads and 20% writes. The workload possesses a skewed access pattern in which, a small set of 10000 keys are 1000 times more likely to be chosen compared to the other keys for writes. This workload is inspired by real-life industrial examples of workload concerning twitter feeds of celebrities. A example scenario is of handling tweets from 10000 celebrities who are more likely, say by a 1000×, to tweet than a non-celebrity user on a particular day, and whose posts on that day are *significantly* (e.g. 8 million times) more likely to be read. For the sake of simplicity, in this example, we assume that the database is stored on a single 16 GB VM, bought by $8000 using AWS, although these observations generalize to more realistic multi-node systems as we will demonstrate later in the paper.

Precise Estimation of I/Os. In this scenario, running the optimization engine with the worst-case cost model results in a configuration $\Omega_{\text{worst-case}}$ as $T = 8$, $K = 7$, $Z = 1$, $L = 3$, $Y = 0$, $M = 16$ GB, $M_B = 3.2$ GB, $M_F = 12.8$ GB, $M_{FP} = 0.29$ GB, $M_{BF} = 12.5$ GB, where the FPR of the design is 0.006. The estimated number of I/Os incurred by $\Omega_{\text{worst-case}}$ according to the worst-case cost model is 40×10^9. However, when the same design is analyzed using the proposed distribution-aware cost model, the resulting I/O cost turns out to be 33×10^9, which is significantly lower. The main reason for the difference in cost is that, the proposed cost model perceives a design very differently from the worst-case model. Given a design, unlike the worst-case model, the proposed model probabilistically estimates and the number of levels and the number of the runs within each level of the design which every query within a workload may touch. The model exploits the characteristic of *early stopping* and takes into account that different queries can touch data residing at different levels of the tree and that result of any query may be returned much earlier than the last level. This leads to the first property of teh proposed model as follows:

Property 1: *Due to early stopping, the estimated I/O cost of a particular design is significantly lower in the distribution-aware cost model as compared to the worst-case model.*

The implication of property 1 is that the performance of a design is significantly underestimated by the worst-case cost model. Therefore, if a user aims to achieve a certain performance for an application, worst-case models are likely to suggest designs that deliver much higher than the targeted performance, but as a result, the cost to be payed (for purchasing storage and compute resources) for supporting such designs also increases, thus significantly overcharging the application. On the other end, if a user has a budget fixed for an application, the worst-case model will select designs with performance much lower than what can be afforded with the cost.

Improving the Search Process Over the Design Space. It evidently follows from property 1 that, owing to the inaccuracies in the model design, searching for designs using worst-case models leads to design suggestions that are highly sub-optimal. Therefore, we incorporated the distribution-aware cost model within the search process to examine the difference in the suggested designs. We observe that the optimal design suggested by the proposed cost model, $\Omega_{\text{distribution-aware}}$, turns out to be $T = 32$, $K = 13$, $Z = 1$, $L = 2$, $Y = 0$, $M = 16$ GB, $M_B = 12.8$ GB, $M_F = 3.2$ GB, $M_{FP} = 0.29$ GB, $M_{BF} = 2.91$ GB, with the FPR being is 0.31. With this design, the I/O estimate according to the distribution-aware model is about 19×10^9 I/Os which shows that the search algorithm selects better designs compared to those of the worst-case models. We pinpoint to the fact that with the same financial budget of \$8000 and the same cloud provider, distribution-aware cost models enables users to select better designs that offer higher performance. This leads to the second property of the model as follows:

Property 2: *The distribution-aware cost model can select better designs that offer lower performance, compared to worst-case models, with the same budget and the same cloud provider.*

The reason behind property 2 is that the distribution-aware lookup cost, accounting for early stopping, seeks to increase the size of the buffer in order to maximize the number of buffer hits of the *special* keys that are accessed more frequently. Because the special keys are likely to live in the buffer due to the workload access pattern, the cost model trades off a larger buffer size with a reduced FPR. On the contrary, the worst-case model does not account for the benefits of increasing the buffer size. Due to this, we observe that the worst-case I/O cost is much higher on the first design than the second design: the worst case I/O cost on the second design is 54×10^9 I/Os.

Generation of Fundamentally Different Designs. We perform a generalized analysis by extending the above analysis to include multiple cloud providers and a wide range of budget from \$500/month to \$50000/month. We include different workloads with varied read percentages in the above case study. For each budget, we generate the best performance using both the worst-case and the distribution-aware model and compute the average of the factor by which the the number of I/Os were reduced with the proposed model over the worst-case model. In Figure 1, the x-axis represents the different workloads and the y-axis shows the times of improvement in performance. Each line in the figure represents a different query distribution imposed on the above example. For example, uniform distribution means all keys from the workload are read and updated with equal probability. A 60% skew means that the point lookups

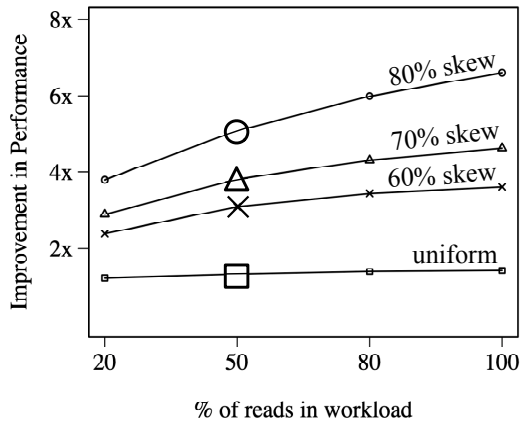

Figure 1: The figure shows the holistic implications on performance that can be achieved within the exhaustive continuum comprising of multiple designs, multiple cloud providers, and different budgets.

targeted the special keys with a probability of 0.6% and the residual keys with a probability of 0.4%. Similarly, the definition holds true for 70% and 80% skews. We observe that, regardless of the distribution of the workload, the proposed cost model significantly improves the performance of read-intensive workloads as the model reasons about the cost of point-lookups differently from that of the worst-case model. We observe that for uniform distribution the distribution-aware cost model performs about 1.5× better but is particularly compelling for skewed workloads in which the performance increases as the skew percentage increases. This is mainly because the proposed model takes into account the possibilities of answering point lookups of frequently accessed keys that need disk I/Os less often. From this observation, we deduce the third property as follows:

Property 3: *The difference in performance created by the distribution-aware cost model with respect to the the worst-case model at the level of designs within a single VM, translates to generating fundamentally different configurations in the continuum that differ in the choices of hardware, combinations of VM instances, or the cloud provider for the same budget.*

With the increase in the growth of workloads and data, the error in the estimation of I/Os for worst-case models significantly increases. In fact, as the workload grows beyond a a certain point, worst-case models turn out to be counterproductive within cloud computing environments. At such stages, I/O estimation becomes all the more crucial as it regulates either the cost payed for a targeted performance or the performance obtained for a fixed budget. To this end, the proposed distribution-aware cost model focuses to closely capture the cost-performance relationship that results in accurate estimation of the I/O cost.

Algorithm 1 Construction of Hardware Space

Output: $H = \{H_{s_i}\}, \forall s_i \in S$
procedure GETALLVMCOMBINATIONS(S)
  H_{s_i} = NULL
  for all $s_i \in S$ **do**
    set $\Lambda =< \lambda_{i,1}, \lambda_{i,2}, \cdots, \lambda_{i,k_i} >$ to $< 0, 0, \cdots, 0 >$
    for machines $\leq$ MAX_CLUSTER_SIZE **do**
      getCombination(s_i, 1, k_i, machines, Λ)
procedure GETCOMBINATIONS(s_i, start, end, machines, Λ)
  if machines == 1 **then**
    for $i \in$ [start .. end] **do**
      $\Lambda[i]+ = 1$; Add configuration Λ to H_{s_i};$\Lambda[i]- = 1$
  else
    Λ[start]$+ = 1$
    getCombination(s_i, start, end, machines-1, Λ)
    Λ[start]$- = 1$
    if start != end **then**
      getCombination(s_i, start+1, end, machines, Λ)

4 CPU

We created an extensive benchmark on RocksDB and WiredTiger to capture CPU utilization overhead and scaling of modern storage engines [13, 27] on varying hardware and workloads to understand the effects of the design features used in the I/O models, as well as what additional features we can use to navigate this tradeoff. In particular, the benchmark varies inputs (dataset size, workload) and I/O design features ($T, K, Z, M_{BF}, M_{FP}, M_B, concurrentthreads$) to isolate effects on the CPU utilization, as well as introducing additional features such as compression. We discovered that most I/O design features have a negligible effect on CPU utilization with the exception of two: a) multi-threading and b) compression.

Benchmark Setup. To capture CPU utilization overhead we compose variations on different workload conditions by running a benchmark consisting of 5 million writes and 50 thousand point queries run at different entropies and with varied compression schemes. This is shown in Table 1, where we display the *additive* CPU overhead and storage space reduction associated with three workload entropy contexts and three different compression algorithms for one hardware configuration. We use machines with Core i5 processors, 16GB DDR4 RAM, and dual drives with 256 GB of 2.5" SATA SSD and 256GB of m.2 SSD (Samsung 960 EVO) with Ubuntu 16.04LTS.

Compression for trading CPU against I/O footprint. Lossless compression algorithms [1, 14, 15, 20] can be deployed before storage block serialization to reduce the I/O cost of any given physical read or write to storage device. These compression algorithms make a trade, utilizing free available CPU in order to realize significant I/O savings.

In an otherwise I/O bound context, utilizing a compression algorithm adds pre-I/O steps of CPU-intensive compression and decompression steps on the data. The nature of this overhead creates an *additive* penalty to latency due to pure CPU-bound steps. On a CPU-overloaded machine, this lightens the overall load on the physical I/O device at the expense of potentially blocking query processing with more intensive CPU load. This means, by choosing more heavyweight compression algorithms, we can further increase the realized storage and I/O savings as long as the bare metal physical machine can handle the additive CPU load. Cosine's CPU model, allows CPU estimation which is more precise than pure Disk Access Models which are only concerned with I/O access costs from block or page movements across the memory and storage hierarchy.

Workload Entropy	Compression Type	Get Overhead (%)	Put Overhead (%)	Space Reduction (%)
0.1	Snappy	0.7	11.06	82.87
	ZLIB	10.25	21.75	91.22
0.25	Snappy	0.53	8.21	68.2
	ZLIB	25.45	31.26	83.18
0.5	Snappy	0.53	12.5	46.86
	ZLIB	26.77	12.1	70.28
0.75	Snappy	0.95	5.17	27.64
	ZLIB	28.65	22.29	59.04

Table 1: A compression behavior lookup table for CPU and storage, given hardware and entropy context.

5 ADDITIONAL DETAILS ON EXPERIMENTAL SETUP

We provide the experimental parameters used for cloud pricing policies in Table 2. These providers, pricings and specifications create a set of hardware choices when determining the best cloud configuration.

6 RELATED WORK

We discuss related areas of multi-node cloud cluster management, cloud resource pricing, utilization and guarantees, and other data management work in cost optimization and their relation to Cosine's focus on optimality in systems and data structure design in the cloud.

Multi-Node Considerations: Sharding and Data Partitioning. Distributed data systems [4, 11, 16, 22–24] have grown in popularity due to the ease of horizontal scaling with cloud deployments to accommodate workload and data growth. In order to scale, these systems shard their data, distributing it among compute nodes and dynamically rebalancing as the total dataset size changes to maintain scaling properties without a priori guarantees on the data's growth rate. Current state-of-the-art systems use workload unaware methods, such as consistent hashing [11, 19, 23, 24], simpler hash-based sharding schemes [4, 22], or sharding by key ranges [16, 22]. These systems intend only to balance overall data size across all compute shards, usually with respect to a weighting determined by workload demands. Workload-aware partitioning has been advocated in a number of systems and schemes for sharding and data placement, in order to improve distributed data layout [2, 5, 6, 9, 12, 17], but is not commonly used in production systems.

Our work can be used A) orthogonally to the sharding scheme, or B) as part of a co-optimization. Cosine can be used to optimize per-node data layout after either workload-unaware or workload-aware sharding, providing strictly additive benefits by finding the best data structure design for each node. In addition, our optimization problem formulation integrates with sharding: we design the distributed partitions of the data layout in a weighted manner that minimizes the criterion function for total cost (often measured as data movement, or over-the-network I/O). Cosine can relax its assumption of proportional sharding by straightforwardly incorporating this cost-model and co-optimizing the sharding scheme with the per-node data layout to minimize the total cost over any number of heterogeneous-sized partitions, allowing Cosine to be workload

Table 2: Experimental parameters used for cloud pricing policies across different cloud providers

AWS			
Instance Type	**Default vCPU**	**Memory (GB)**	**Hourly Rate ($)**
r5d.large	2	16	0.091
r5d.xlarge	4	32	0.182
r5d.2xlarge	8	64	0.364
r5d.4xlarge	16	128	0.727
r5d.12xlarge	48	384	2.181
r5d.24xlarge	96	768	4.362
r5d.metal	96	768	4.362

GCP			
Instance Type	**Default vCPU**	**Memory (GB)**	**Hourly Rate ($)**
n1-highmem-2	2	13	0.0745
n1-highmem-4	4	26	0.1491
n1-highmem-8	8	52	0.2981
n1-highmem-16	16	104	0.5962
n1-highmem-32	32	208	1.1924
n1-highmem-64	64	416	2.3849
n1-highmem-96	96	624	3.5773

AZURE			
Instance Type	**Default vCPU**	**Memory (GB)**	**Hourly Rate ($)**
E2 v3	2	16	0.0782
E4 v3	4	32	0.1564
E8 v3	8	64	0.3128
E16 v3	16	128	0.6256
E20 v3	20	160	0.7409
E32 v3	32	256	1.2512
E64 v3	64	512	2.5024

aware. This approach can be extended to consider heterogenous per-node hardware choices across the cluster at the expense of more search time, tapping into a larger space of cluster designs which captures state-of-the-art sharding.

Cloud Resource Pricing, Utilization and Guarantees. Cloud tenants and providers are increasingly sensitive to the price charged for resources (VM compute, network bandwidth, etc) [7, 18, 26]. Cloud providers focus on increasing utilization and revenue by balancing the tension between ensuring the availability of extra resources to guarantee performance thresholds, and selling these resources (e.g. bandwidth, or CPU credits available) for additional revenue [21].

While the monetary dimension is a concern in this subarea, we are unaware of attempts to forecast and derive the tradeoff between end economic cost and performance of individual application instances in relation to their starting configurations. Cosine is orthogonal to vendors' designs of pricing schemes and can integrate with additional pricing schemes as they arise. While pricing and resource markets can change for monetary costs dynamically, Cosine allows the usage of arbitrary pricing schemes on each cloud provider. Cosine navigates between multiple pricing schemes (or markets) across different providers to satisfy arbitrary balances of cost and performance and provides the cloud tenant the combination of hardware choice and data layout that gives the best performing configuration for their given workoad and preferences.

Data Management as Interpretable Optimization. Beyond the NoSQL context, in data warehousing, the topics of optimal index selection, query planning, and algorithms design are domains with established classic meta problems of search, ranking, selection, and synthesis of efficient configurations, designs and plans (e.g. physical storage, indexes, query planing) [8, 25]. Recently, work using black box machine learning (ML) driven approaches such as neural networks have grown in popularity. The learned approach tries to capture the mappings of each operational context to its best configuration for optimal performance. However, this requires accurate modeling of the holistic system behavior and understanding the interaction of multiple factors to achieve robust performance inference. Any predicted end performance needs to be learned from features and patterns that the ML model extracts from the observed performance traces of the data system [3, 10]. As such, generalizing performance prediction across differing hardwares, workloads, and other contexts is challenging.

Cosine's recommendations are more robust than pure ML approaches. It uses explicit I/O and CPU models which break down end performance prediction into clearly delineated parts, which black box approaches like neural networks have no guarantee of when predicting and choosing configurations. This also allows us to gain insights into the configurations and improve generalizability to new hybrid designs. Cosine is responsive - it does not need a runway of training time for trying configurations or tuning of ML model parameters, which allows rapid deployment and cost reduction by reducing the need for full runs on test environments that learned models need to accumulate traces for training. Additionally, in what-if mode, Cosine has the potential to reason about "neighboring" configurations and suggest an alternate configuration alongside the optimal, where a relatively small increase in cost beyond the user's maximum cost threshold could still result in a favorable performance gain.

REFERENCES

[1] 2019. LZ4 - Extremely Fast Compression algorithm. (2019).

[2] Sharad Agarwal, John Dunagan, Navendu Jain, Stefan Saroiu, Alec Wolman, and Harbinder Bhogan. 2010. Volley: Automated Data Placement for Geo-distributed Cloud Services. In *Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation (NSDI'10)*. USENIX Association, Berkeley, CA, USA, 2–2.

[3] Dana Van Aken, Andrew Pavlo, Geoffrey J Gordon, and Bohan Zhang. 2017. Automatic Database Management System Tuning Through Large-scale Machine Learning. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 1009–1024.

[4] Apache. 2019. Cassandra. *http://cassandra.apache.org* (2019).

[5] Aaron Archer, Kevin Aydin, Mohammad Hossein Bateni, Vahab Mirrokni, Aaron Schild, Ray Yang, and Richard Zhuang. 2019. Cache-aware Load Balancing of Data Center Applications. *Proc. VLDB Endow.* 12, 6 (Feb. 2019), 709–723.

[6] Nicholas Ball and Peter Pietzuch. 2013. Skyler: Dynamic, Workload-Aware Data Sharding across Multiple Data Centres. (2013).

[7] Hitesh Ballani, Paolo Costa, Thomas Karagiannis, and Ant Rowstron. 2011. The Price Is Right: Towards Location-Independent Costs in Datacenters. ACM HotNets.

[8] Surajit Chaudhuri. 1998. An Overview of Query Optimization in Relational Systems. In *Proceedings of the ACM Symposium on Principles of Database Systems (PODS)*. 34–43.

[9] Carlo Curino, Yang Zhang, Evan P C Jones, and Samuel Madden. 2010. Schism: A Workload-Driven Approach to Database Replication and Partitioning. *Proceedings of the VLDB Endowment* 3, 1 (2010), 48–57.

[10] Sudipto Das, Miroslav Grbic, Igor Ilic, Isidora Jovandic, Andrija Jovanovic, Vivek R. Narasayya, Miodrag Radulovic, Maja Stikic, Gaoxiang Xu, and Surajit Chaudhuri. 2019. Automatically Indexing Millions of Databases in Microsoft Azure SQL

Database. In *Proceedings of the 2019 International Conference on Management of Data (SIGMOD '19)*. ACM, New York, NY, USA, 666–679.

[11] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. 2007. Dynamo: Amazon's Highly Available Key-value Store. *ACM SIGOPS Operating Systems Review* 41, 6 (2007), 205–220.

[12] Diego Didona and Willy Zwaenepoel. 2019. Size-aware Sharding For Improving Tail Latencies in In-memory Key-value Stores. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*. USENIX Association, Boston, MA, 79–94.

[13] Facebook. 2019. RocksDB. *https://github.com/facebook/rocksdb* (2019).

[14] Facebook. 2019. Zstandard - Fast real-time compression algorithm. (2019).

[15] Google. 2019. Snappy - A Fast Compressor/Decompressor. (2019).

[16] Bram Gruneir. 2017. Scalable SQL Made Easy: How CockroachDB Automates Operations .

[17] Monika Henzinger, Stefan Neumann, and Stefan Schmid. 2019. Efficient Distributed Workload (Re-)Embedding. *Proc. ACM Meas. Anal. Comput. Syst.* 3, 1, Article 13 (March 2019), 38 pages.

[18] Darrell Hoy, Nicole Immorlica, and Brendan Lucier. 2016. On-Demand or Spot? Selling the Cloud to Risk-Averse Customers. In *Proceedings of the 12th International Conference on Web and Internet Economics - Volume 10123 (WINE 2016)*. Springer-Verlag New York, Inc., New York, NY, USA, 73–86.

[19] David Karger, Eric Lehman, Tom Leighton, Rina Panigrahy, Matthew Levine, and Daniel Lewin. 1997. Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web. In *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing (STOC '97)*. ACM, New York, NY, USA, 654–663.

[20] Jean loup Gailly and Mark Adler. 2017. Zlib - A Massively Spiffy Yet Delicately Unobtrusive Compression Library. (2017).

[21] Jeffrey C. Mogul and Lucian Popa. 2012. What We Talk About when We Talk About Cloud Network Performance. *SIGCOMM Comput. Commun. Rev.* 42, 5 (Sept. 2012), 44–48.

[22] MongoDB. 2019. Online reference. *http://www.mongodb.com/* (2019).

[23] Oracle. 2018. Introducing Oracle Database 18c. *White Paper* (2018).

[24] OrientDB. 2019. OrientDB Documentation: Auto Sharding Index Algorithm.

[25] Patricia G. Selinger, Morton M. Astrahan, Donald D. Chamberlin, Raymond A. Lorie, and Thomas G. Price. 1979. Access Path Selection in a Relational Database Management System. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 23–34.

[26] Duong Tung Nguyen, Long Bao Le, and Vijay Bhargava. 2018. Price-based Resource Allocation for Edge Computing: A Market Equilibrium Approach. *IEEE Transactions on Cloud Computing* PP (06 2018), 1–1.

[27] WiredTiger. 2019. Source Code. *https://github.com/wiredtiger/wiredtiger* (2019).