- Locality Sensitive Hashing(LSH)
  - Common technique in data clustering, nearest neighbor problem and high dimension data indexing.
  - Use hash function h(x) and combination of several hash functions to make sure similar data have larger possibility to be in the same bucket after hashing.

- p-stable distribution hashing
  - One kind of LSH function that is suitable for numerical data.

Given a point v in d dimension

$$h_{a,b}(v) = \left\lfloor \frac{a.v + b}{w} \right\rfloor$$

$$h_{a,b}(v) : \mathbb{R}^d \to \mathbb{Z}$$

Each hash function is a random hyperplane in d dimensional space where a is generated from p-stable distribution, the hash value is the distance between v and hyperplane.

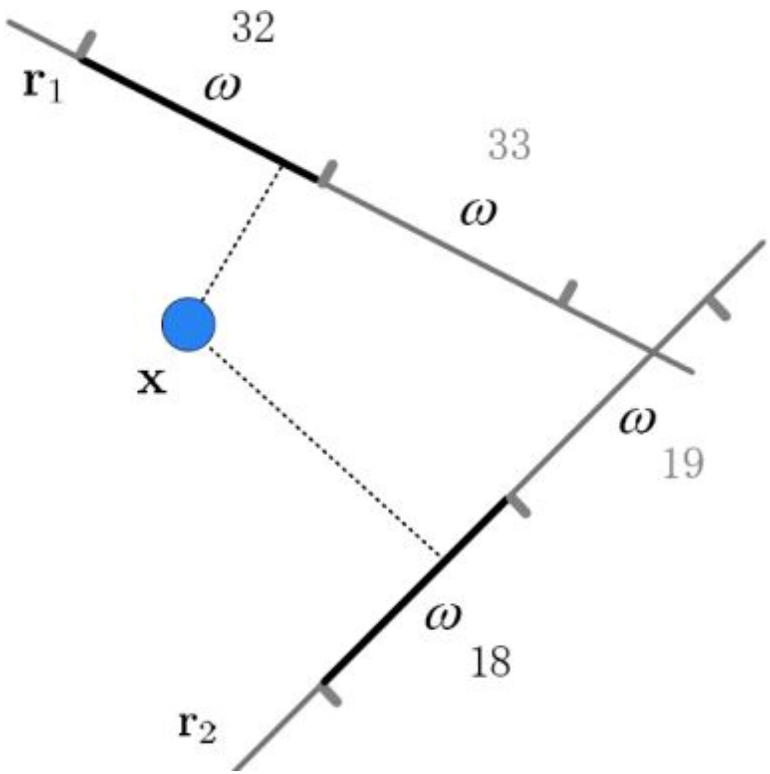
- We cascade hash function for k times to increase accuracy

$$g(v) = (h_1(v), \ldots, h_k(v))$$

An example of LSH in 2D with two subfunctions.

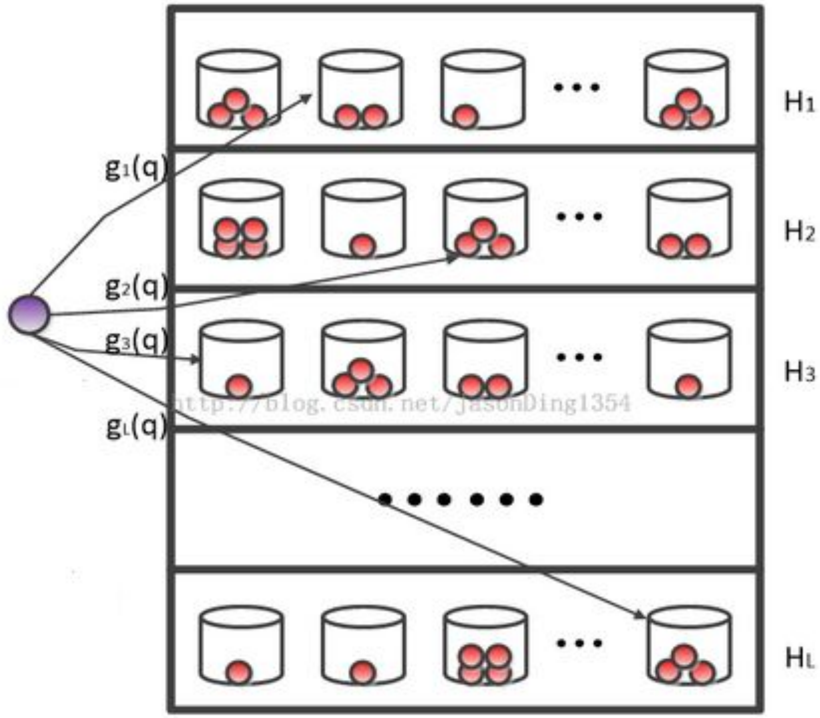
Points close to x have higher possibility to have same hash value as x.

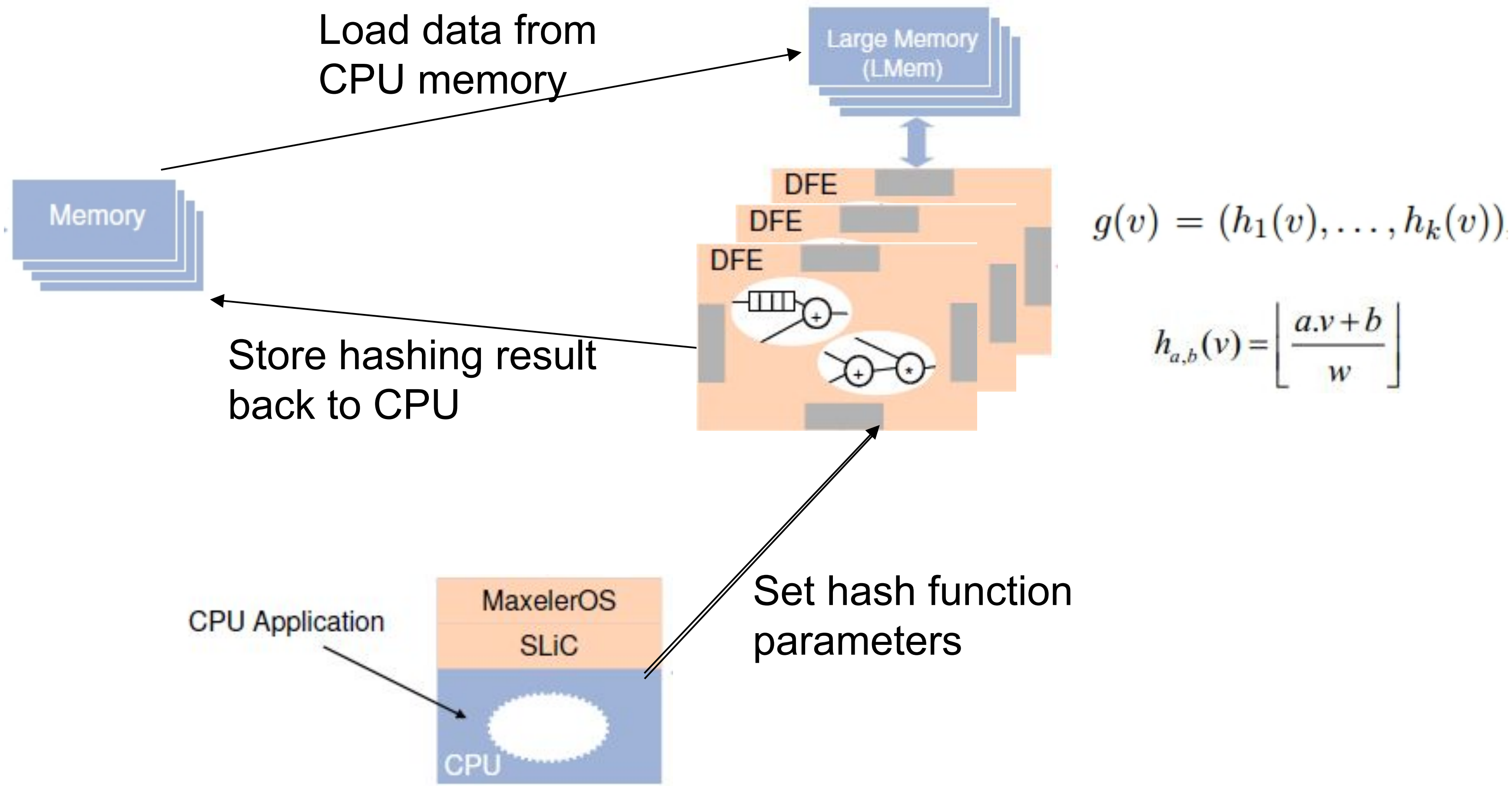
Use LSH to construct an approximate neighbor for x.

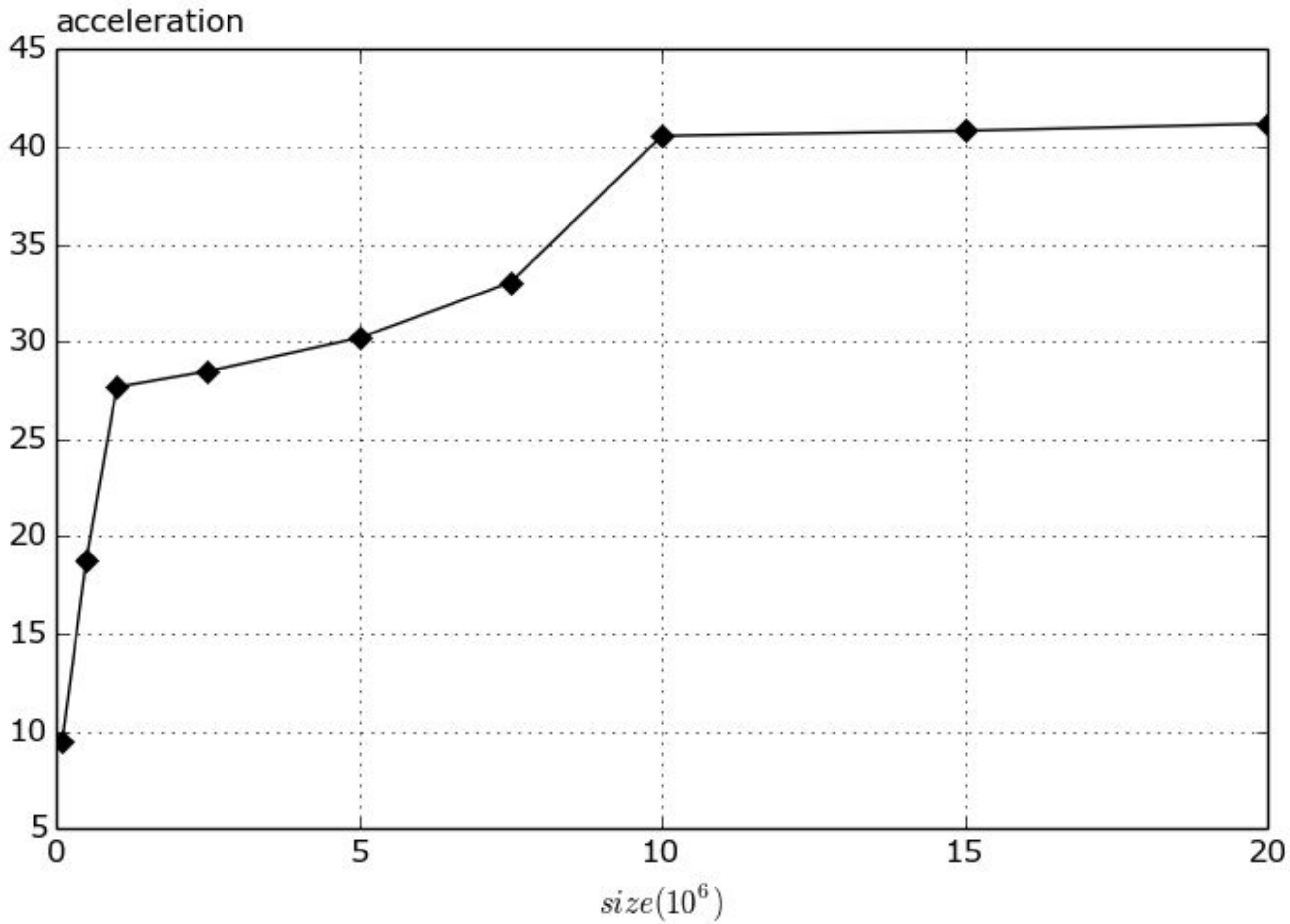


- To increase the recall rate, we still have to repreat function g for L times.

- The whole design is very suitable for dataflow engine acceleration.
  - Each $g_i(q)$ is independent
  - Each $h_j(q)$ in g is independent
  - Data point is independent




Load data from CPU memory
Large Memory (LMem)
Memory
DFE
DFE
DFE
$g(v) = (h_1(v), \ldots, h_k(v))$
$h_{a,b}(v) = \left\lfloor \frac{a.v+b}{w} \right\rfloor$
Store hashing result back to CPU
CPU Application
MaxelerOS
SLiC
CPU
Set hash function parameters



- 2D input data with 16 cascade $h_j(q)$ functions
- FPGA: only use one kernel
- CPU: one core, C++ code compiled with icpc with -O3
- Maximum acceleration is about 42 times

- Acceleration can be roughly estimated by this:

$$h_{a,b}(v) = \left\lfloor \frac{a.v+b}{w} \right\rfloor$$

$$g(v) = (h_1(v), \ldots, h_k(v))$$

- D dimension data with K level cascade
- One function g contains:
  - D*K multiplies
  - K addition
  - K division
- The maximum acceleration should be related to D*K
- If we repeat the kernels, calculate many function g at the same time, there will be more acceleration.
- LSH with dataflow engine will be a perfect choice for spatial data indexing