Revisiting POPCOUNT Operations in CPUs/GPUs

Chenfan Sun - chenfs@cs.washington.edu

Motivation

POPCOUNT:

1	0	0	1	0	1

↓

3

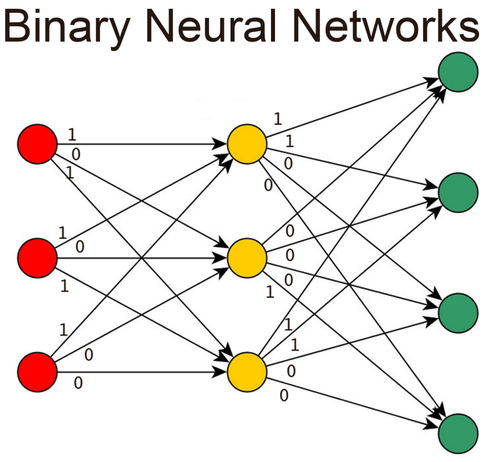
Hamming Distance[1]

$$D_H = \sum_{i=1}^{k} | x_i - y_i |$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

Binary Neural Networks [2]



Approach

Data Set Size: 2MB (Fits in Cache), 1GB

CPUs

Vector Extension { SSSE3, AVX-2 } + Algorithm { Harley-Seal, Swar, Lookup }

GPUs

Data Width { 32 bits, 64 bits, 128 bits } + Algorithm { CUDA popcnt, Swar, Harley-Seal } + Reduction { CPU Reduction, GPU Reduction, GPU Shuffle }

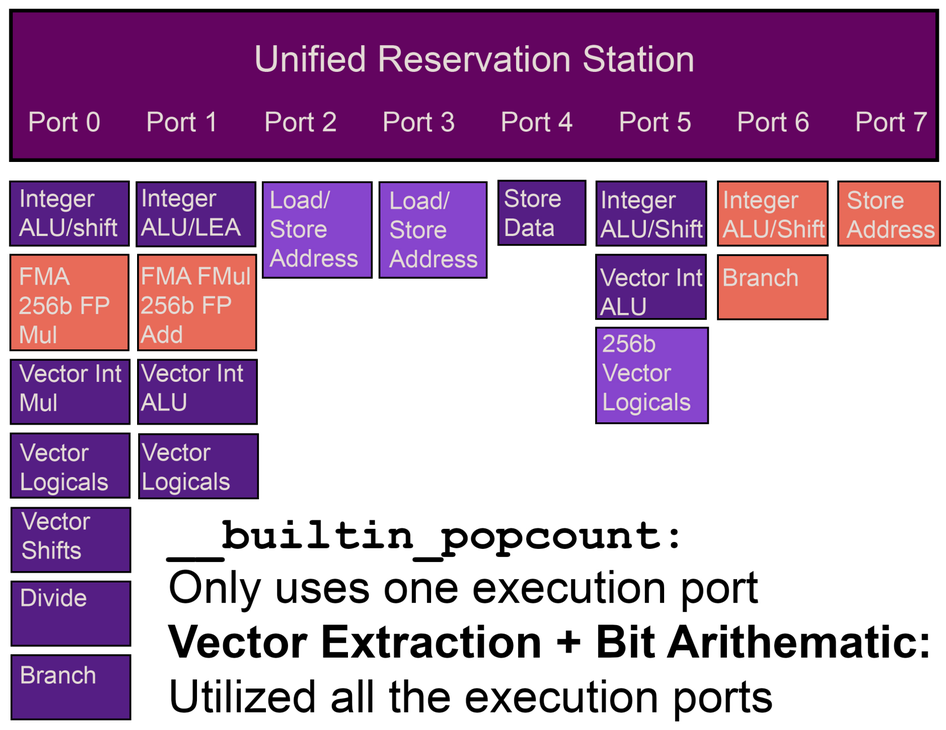
```
a = 0b1101                  # initial number
a0 = (a >> 0) & 0b0101      # every other bits
a1 = (a >> 1) & 0b0101      # remaining bits
b = a0 + a1                 # 1s in 2-bit slice
b0 = (b >> 0) & 0b0011      # = 0b0001
b1 = (b >> 2) & 0b0011      # = 0b0010
ans = b0 + b1               # 1s in 4-bit slice,
ans: 3                      # final answer in 2-bit word
```

Testbed

Intel Haswell i7-4790K
GeForce GTX Titan X (Maxwell)
DDR3 1333MHz Dual-Channel

Background

Intel Haswell Execution Ports



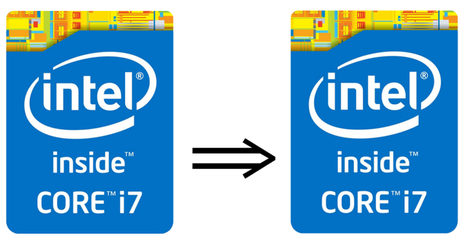
`__builtin_popcount:`
Only uses one execution port

Vector Extraction + Bit Arithematic:
Utilized all the execution ports

Conclusion

Better Performance for *Compute Bound*:

More Powerful Processer



Haswell i7-4790K ⇒ Haswell i7-6700K

Wider Vector Extension

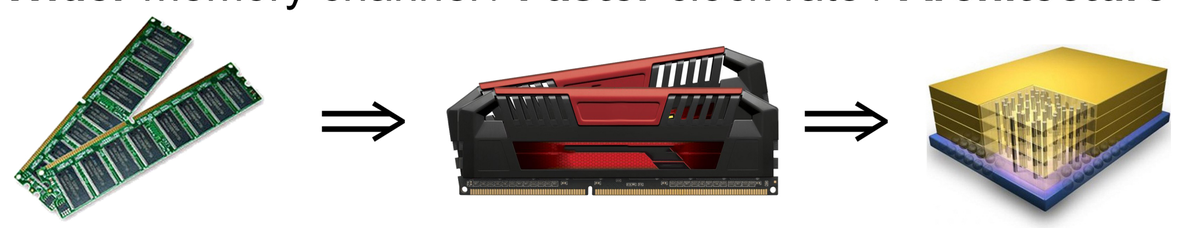
{ SSSE3, AVX-2 } ⇒ { AVX-512 }

Better Algorithm

{ Lookup Table } ⇒ { Harley-Seal }

Better Performance for *Memory Bound*:

Wider memory channel / **Faster** clock rate / **Architecture**

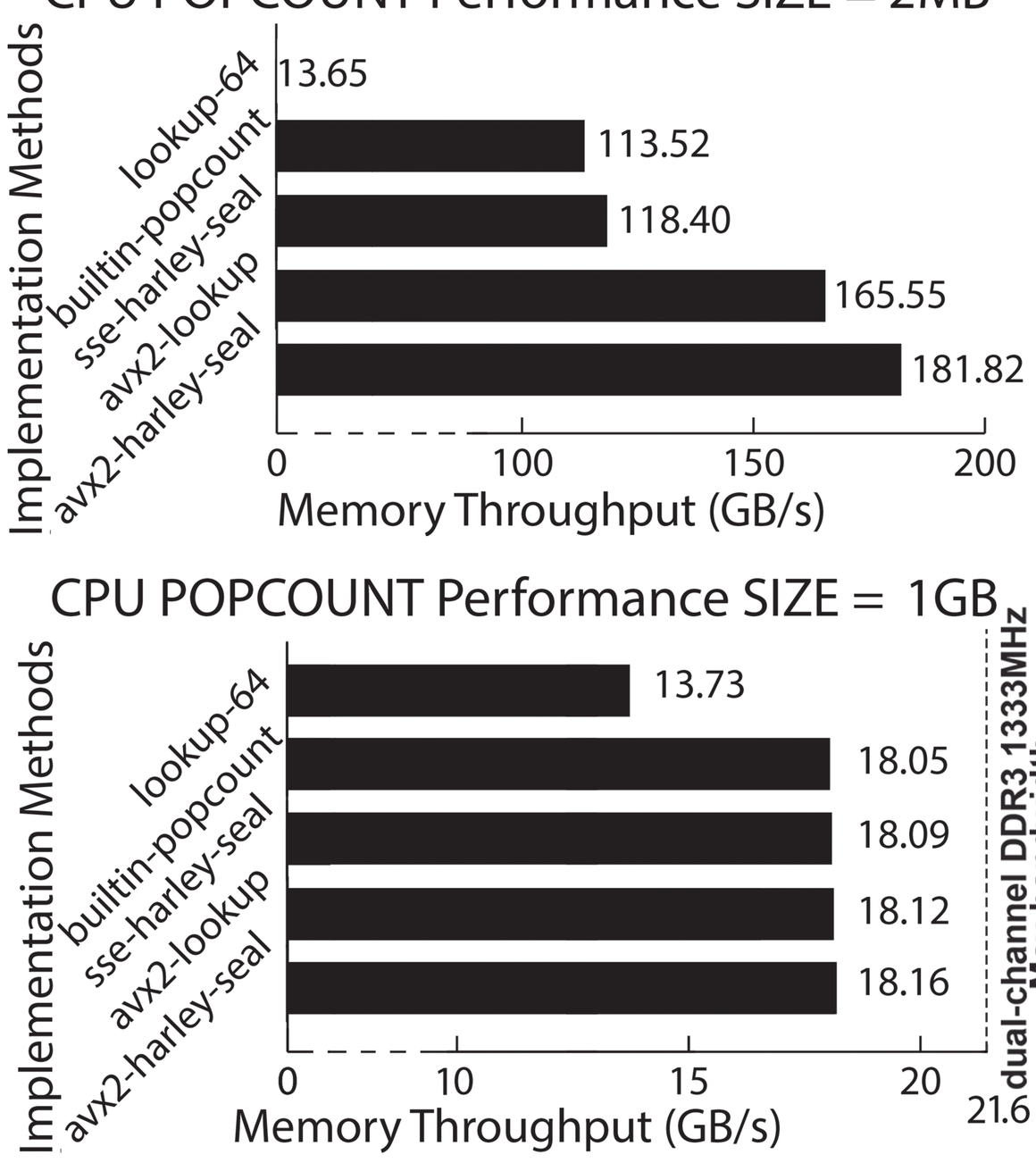


Dual-Channel DDR3 1333MHz ⇒ Qual-Channel DDR4 2133MHz ⇒ Die-Stack Memory

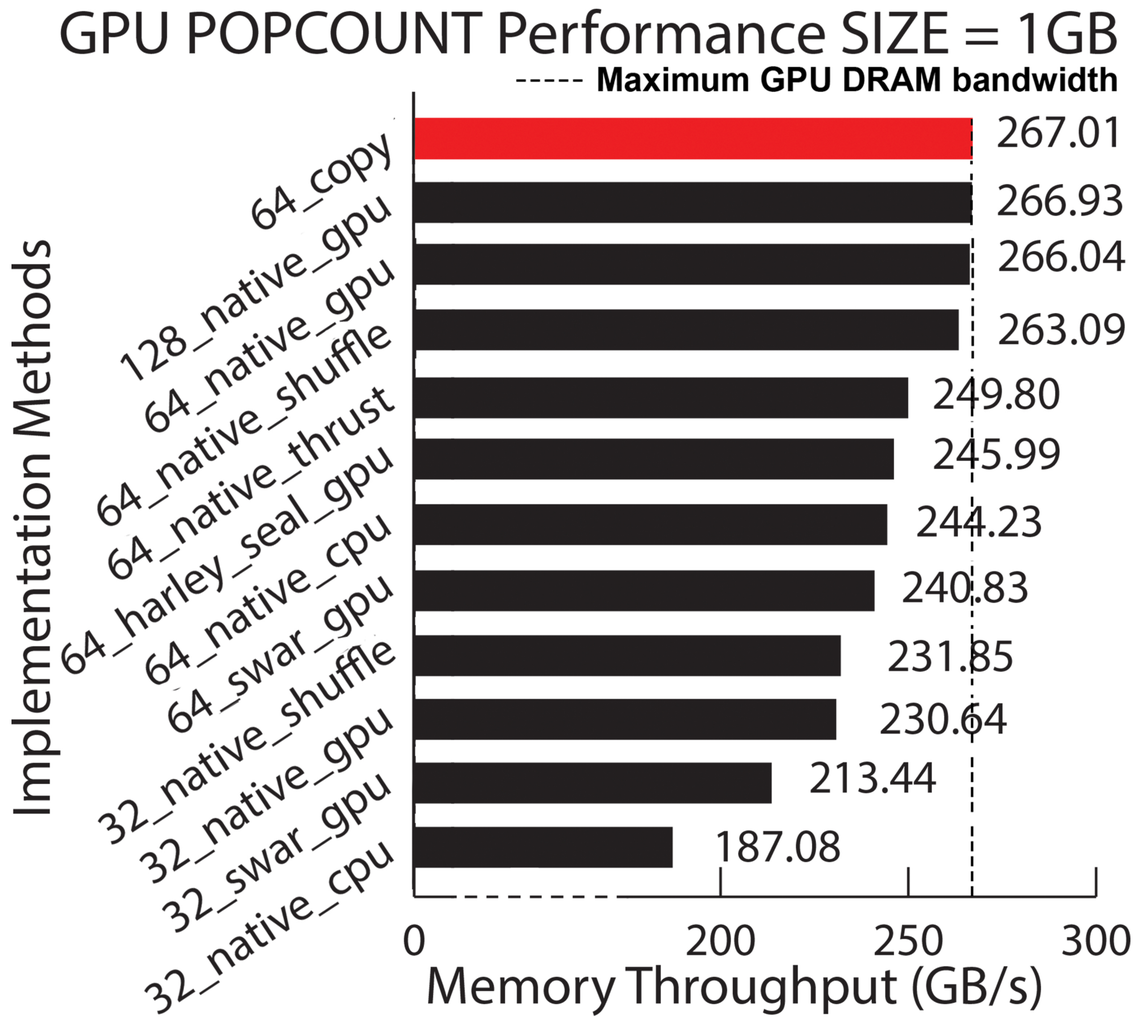
Algorithmic choices do not matter for memory bound.

Results

CPUs



GPUs



[1] V.T. Lee et al., "NCAM: Near-Data Processing for Nearest Neighbor Search," in *arXiv 2016*.
[2] M. Rastegari et al., "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks," in *ECCV 2016*, Amsterdam, Netherlands.