GPU H/W and S/W

Hyesoon Kim

Georgia Tech

Arithmetic Intensity

$$\text{Arithmetic Intensity (AI)} = \frac{\text{Number of floating-point operations (FLOPs)}}{\text{Bytes of memory transferred}}$$

Code example: CUDA K-mean

```
__global__ void assignClusters(const float *points, const float *centroids, int *assigı
    int idx = blockIdx.x * blockDim.x + threadIdx.x;
    if (idx < n_points) {
        int best_cluster = 0;
        float min_dist = FLT_MAX;

        for (int c = 0; c < k; c++) {
            float dist = 0.0f;
            for (int d = 0; d < dim; d++) {
                float diff = points[idx * dim + d] - centroids[c * dim + d];
                dist += diff * diff; // 1 FLOP multiply + 1 FLOP add per dimension
            }
            if (dist < min_dist) {
                min_dist = dist;
                best_cluster = c;
            }
        }
        assignments[idx] = best_cluster;
    }
}
```

FLOPS = dim x (1 sub + 1 mul + 1 add) = dim x 3
FLOPS per point = k clusters x dim x 3
Total FLOPS. = n * K * dim * 3

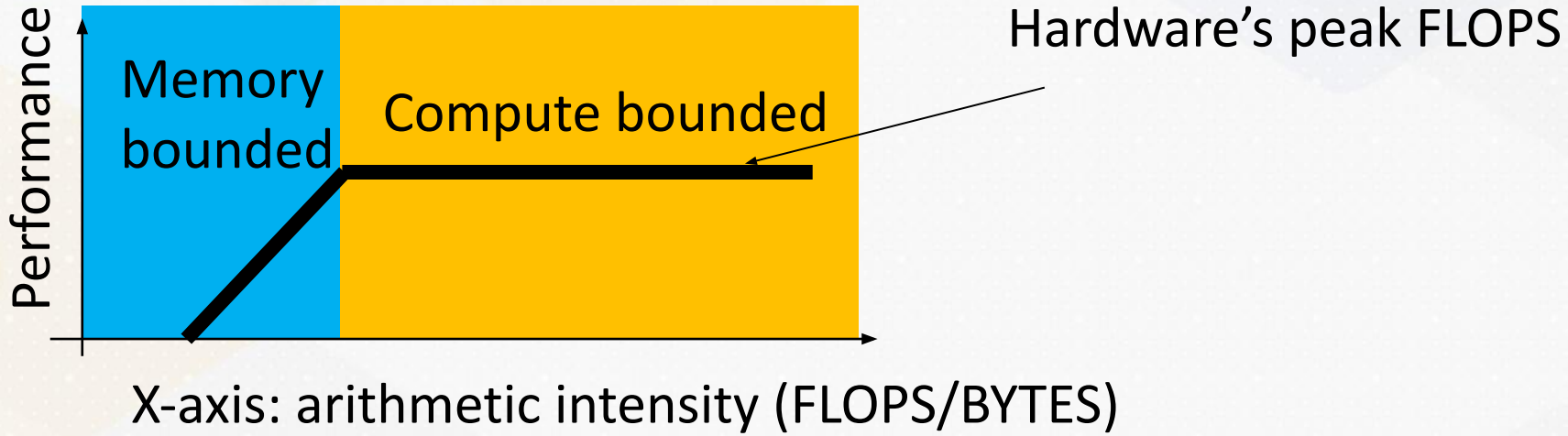
Memory load:
Point data loads: n*dim floats
Centroid data loads: k*dim floats
Store: n
= n *dim * 4 + n*4

- Can you find Arithmetic Intensity?

Code examples: (Perplexity AI, 2025)

Georgia Tech

Roofline Model



- A Visual performance model to determine whether an application (or a processor) is limited by the compute bandwidth or memory bandwidth

Williams, S., Waterman, A., & Patterson, D. (2009). Roofline: An insightful visual performance model for multicore architectures. *Communications of the ACM, 52*(4), 65–76. https://doi.org/10.1145/1498765.1498785

Georgia Tech

What to count for Memory Bytes

- Roofline model: use theoretical minimal memory bytes
- What if we count based on transactional size?
- Based on what to bring from memory or caches

```
#define N 4096

__global__ void matAdd(float *A, float *B, float *C) {
    int idx = blockIdx.x * blockDim.x + threadIdx.x;
    if (idx < N)
        C[idx] = A[idx] + B[idx];
}
```

```
#define N 4096
#define STRIDE 2

__global__ void matAdd(float *A, float *B, float *C) {
    int idx = blockIdx.x * blockDim.x + threadIdx.x;
    if (idx < N/STRIDE)
        C[idx*STRIDE] = A[idx*STRIDE] + B[idx*STRIDE];
}
```

Code examples: (Perplexity AI, 2025)

Georgia Tech

Transactional size

- Between cache: cache block size
- From memory: minimum memory transaction size
- Significant performance impats