Lower Bound on Communication for Matrix Multiplication
• We have a CPU, cache of size M, and large memory
• We want to multiply $n \times n$ matrices $C = A \times B$, which are too big to fit in cache
• Moving data between memory and cache is expensive
• Is there a lower bound on how many reads and writes (moving words between cache and memory) are needed to perform $C = A \times B$?
• Is there an optimal algorithm that attains this lower bound?
Results

- Thm (Hong, Kung, 81) A lower bound on the number of reads and writes is $\Omega(n^3/\sqrt{M})$
- This lower bound is attainable by "tiling" the 3 nested loops (working on square submatrices of A, B and C that all fit in cache simultaneously)
- Both results can be extended to
  - more general code that looks like nested loops accessing arrays (more linear algebra, tensors, CNNs, ...)
  - Memory hierarchies
  - Moving data between processor on a network
- See bebop.cs.berkeley.edu for more details
Lower Bound Proof Sketch

- Inner loop of matmul: $C(i,j) += A(i,k) * B(k,j)$
- Performing one inner loop iteration requires 3 words be in cache
- If I can only fit $M$ words in cache, how many iterations can I do?
- Hard part (next slide): find an upper bound $F$ on the number of iterations I can do
- Need to do $n^3$ iterations => need to refill cache $n^3 / F$ times => #words moved $\geq \left(\frac{n^3}{F}\right) M$
Model iterations over \((i,j,k)\) as an \(n \times n \times n\) cube

If we have at most \(M\) “A squares”, “B squares”, and “C squares” on faces, how many cubes can we have?
If I only have M squares, how many cubes can I “cover”?

# cubes in black box with side lengths x, y and z
= Volume of black box
= x·y·z
= (xz · zy · yx)\(^{1/2}\)
= (\#A\(\square\)s · \#B\(\square\)s · \#C\(\square\)s)\(^{1/2}\)
If I only have M squares, how many cubes can I “cover”?

# cubes in black box with side lengths x, y and z
= Volume of black box
= x·y·z
= (xz · zy · yx)\(^{1/2}\)
= (#A□s · #B□s · #C□s \(^{1/2}\))

(i,k) is in A shadow if (i,j,k) in 3D set
(j,k) is in B shadow if (i,j,k) in 3D set
(i,j) is in C shadow if (i,j,k) in 3D set

Thm (Loomis & Whitney, 1949)
# cubes in 3D set = Volume of 3D set
≤ (area(A shadow) · area(B shadow) · area(C shadow))\(^{1/2}\)
Finishing the lower bound proof

- $F =$ bound on # of loop iterations with $M$ words
- $= \text{bound on #cubes with shadows of size } M$
- $\leq \left( \#\text{entries}_A \times \#\text{entries}_B \times \#\text{entries}_C \right)^{1/2}$
- $\leq \left( M \times M \times M \right)^{1/2} = M^{3/2}$
- #words moved $\geq \left( n^3 / F \right) M = n^3 / \sqrt{M}$