GPU memory

Final project abstract + hypothesis instructions posted (linked right above schedule)

Hardware-aware SPMD

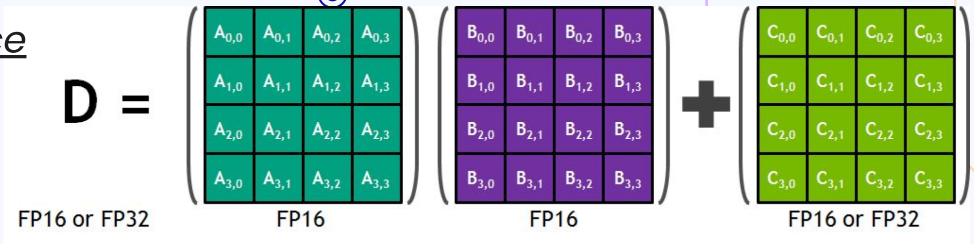
SPMD promises us the power of writing scalar programs and running them with high parallelism

It's (relatively) easy to write a SPMD program that runs, BUT:

- Hardware is going to affect performance
- Same themes we've been talking about (control hazards, data hazards, getting around memory latency) come up in the optimization guide for CUDA

[image source](#)

Tensor cores

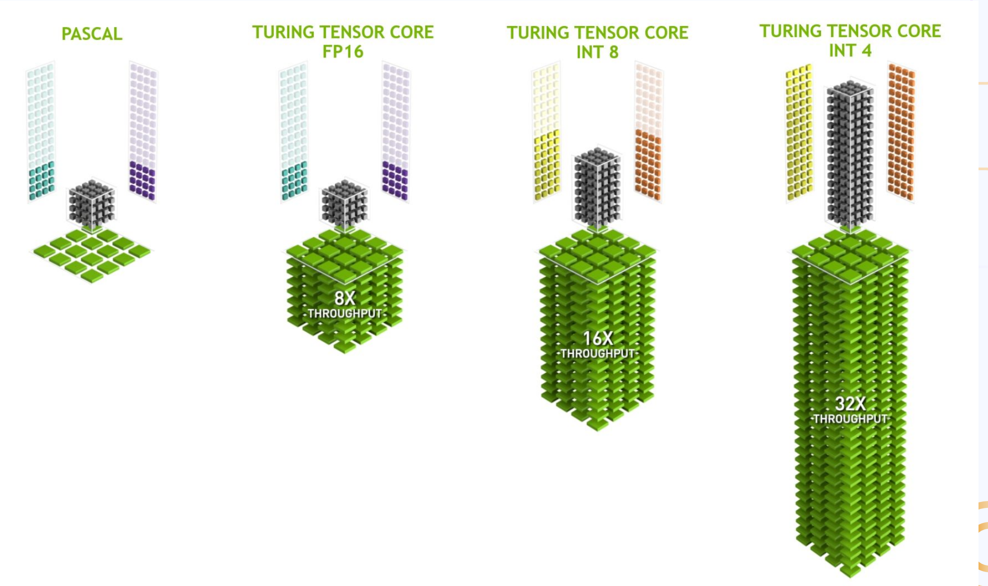


An accelerator on the accelerator (see also Ray Tracing cores)

Circuits on GPUs optimized for AI (matrix operation D = A * B + C)

Uses mixed-precision to speed up math, reduce memory demands

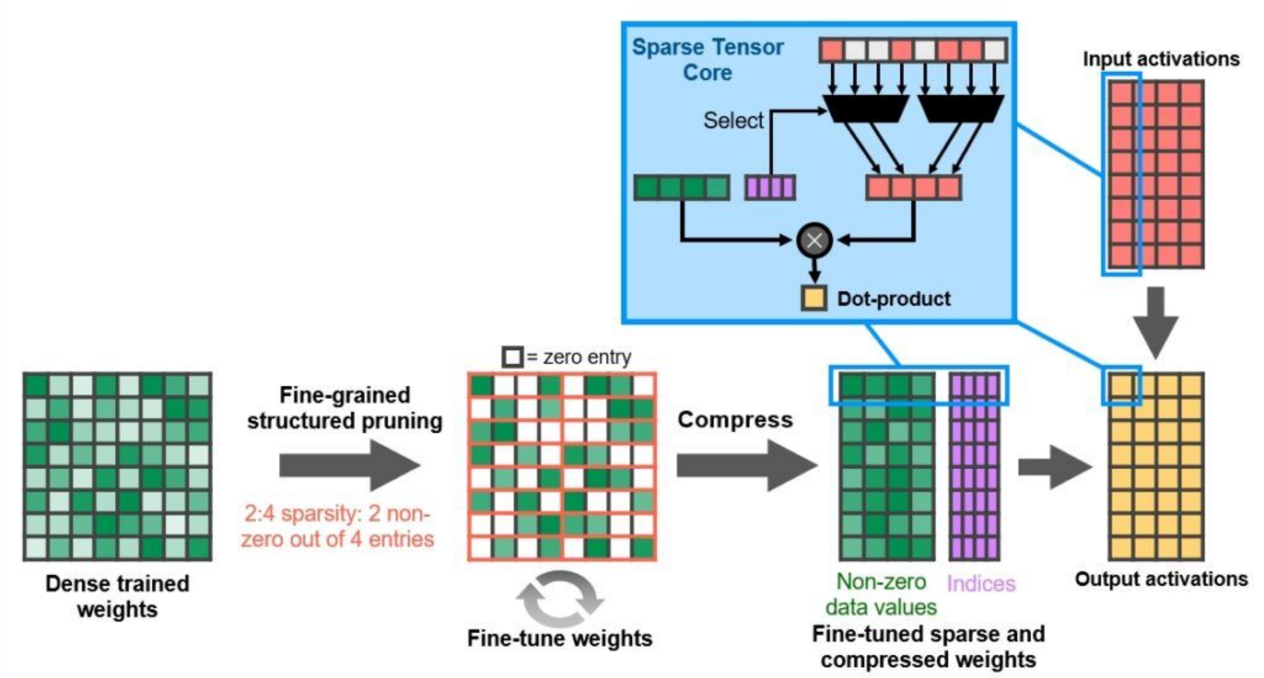
Claim huge (8x-32x per generation) performance gains over SM matrix multiply



[image source](#)

Sparse acceleration

source



source

GA100 SM in Ampere

GPU memory model

Local memory for th[...]
state

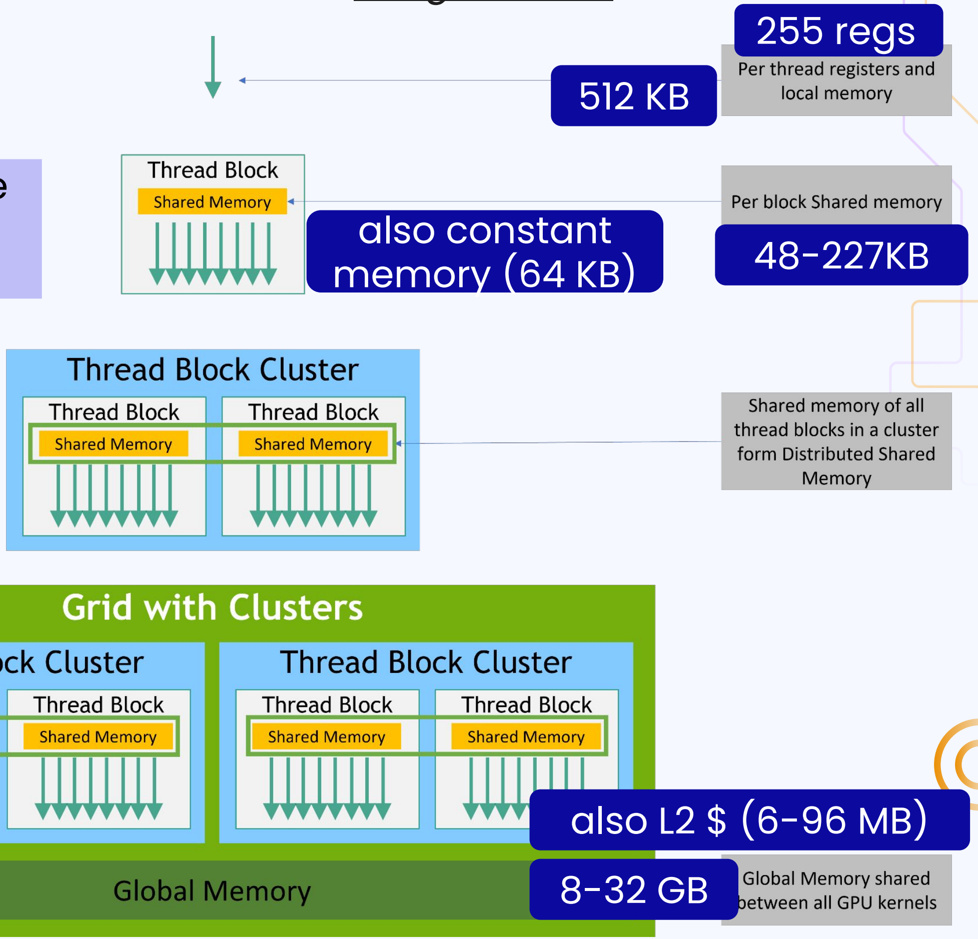
Resides in device memory, not on chip… why?

Shared memory to communicate across threads

How?

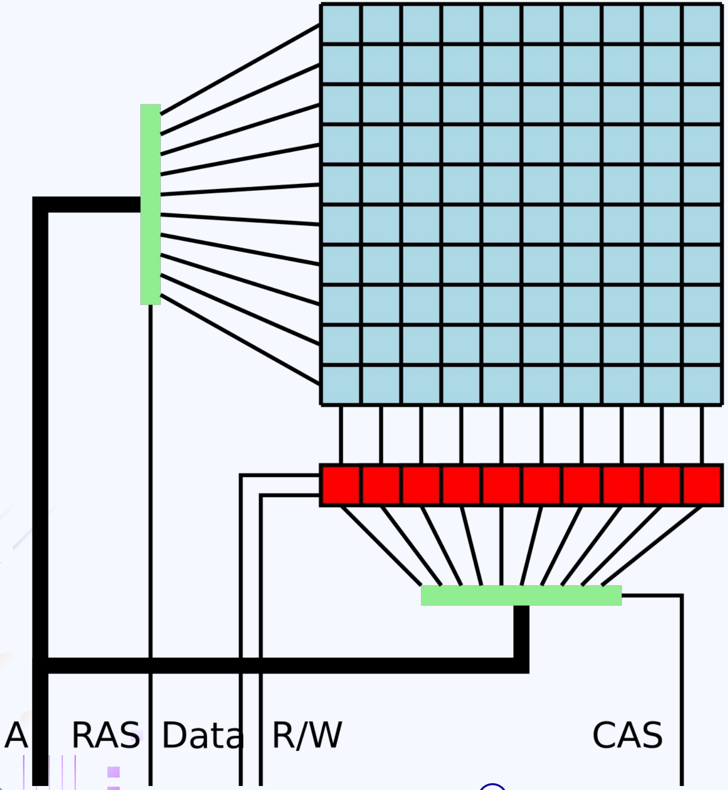
```
__shared__ bool isDone;
…
if (threadIdx.x == 0)
     if (ecc) isDone = true;
…
__syncthreads();
if (isDone) {
     …
}
```

image source



DRAM and GDDR

image source



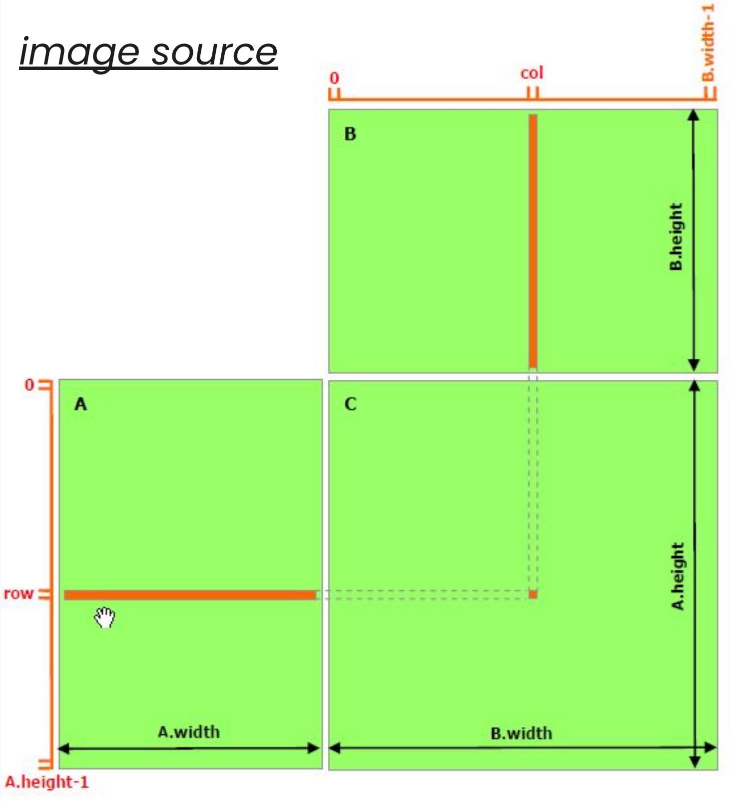
(Not pictured: further organized into banks)
Activating a row takes several cycles
After row is activated, data can be read

Latency/locality tradeoff:
Controller waits for enough requests to a single row before servicing all of them at once

Generally, GDDR (graphics DDR) variants have higher latency and higher bandwidth

GPU memory: matrix multiply

image source



What's wrong with this?

```
//2d blocks of size BLOCK_SIZE * BLOCK_SIZE
int row = threadIdx.y;
int col = threadIdx.x;

float c = 0;
for (int i = 0; i < n; i++) {
    c += GetElement(A, row, i) * GetElement(B, i, col);
}
SetElement(C, row, col, c);
```

Tiling

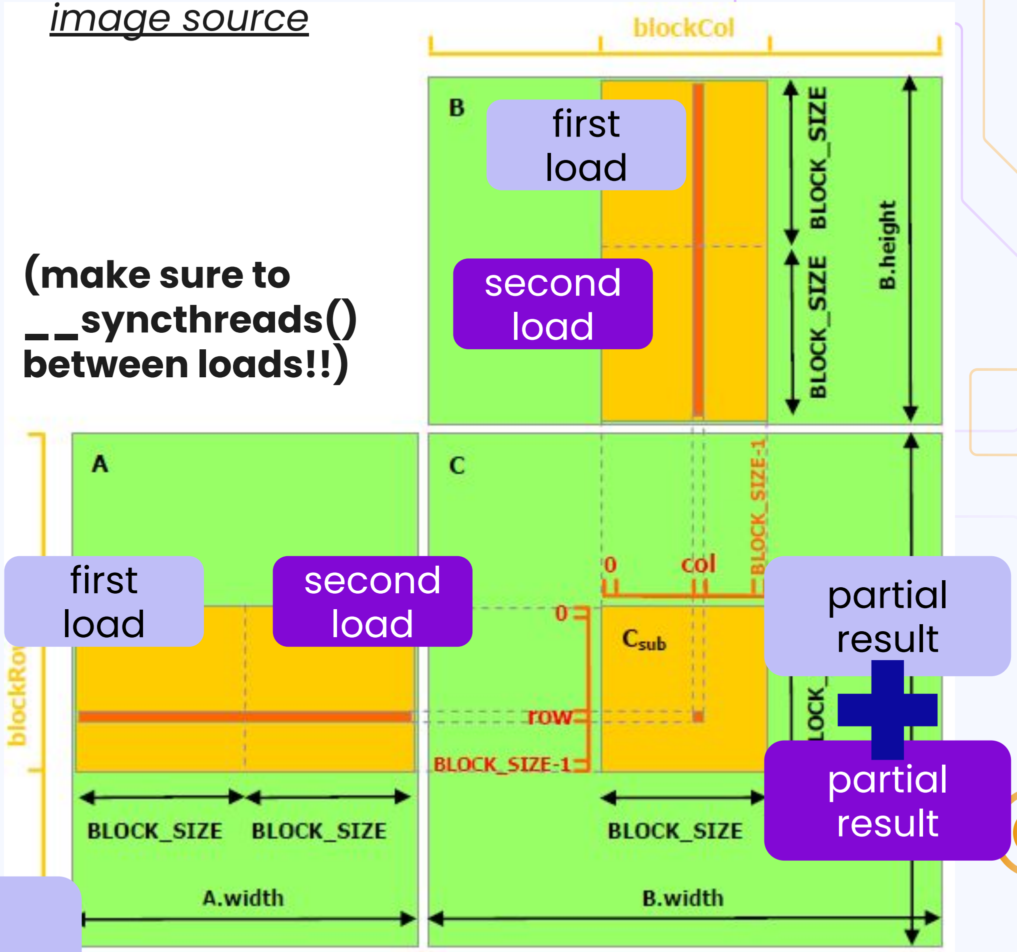
(full code, including loading from CPU memory to device memory, in image source link)

```
//2d blocks of size BLOCK_SIZE * BLOCK_SIZE
int blockRow = blockIdx.y;
int blockCol = blockIdx.x;
int row = threadIdx.y;
int col = threadIdx.x;
...
// for-loop on m (number of tiles to load):
Matrix Asub = GetSubMatrix(A, blockRow, m);
Matrix Bsub = GetSubMatrix(B, m, blockCol);
__shared__ float As[BLOCK_SIZE][BLOCK_SIZE];
__shared__ float Bs[BLOCK_SIZE][BLOCK_SIZE];
As[row][col] = GetElement(Asub, row, col);
Bs[row][col] = GetElement(Bsub, row, col);
```

Each thread does two loads/stores here
Can hardware design make this more efficient?

image source

(make sure to __synchthreads() between loads!!)



Coalesced memory access

Coalescing unit detects if accesses from same warp are in adjacent addresses and performs single, wide access (reduces uses of DRAM line)

Works for both global memory and local memory!

Important for programmer to be mindful of memory indexing

Example: avoid having each thread do its own malloc (source)

```
__shared__ int* data;
if (threadIdx.x == 0) {
    size_t size = blockDim.x * 64;
    data = (int*)malloc(size);
}
__syncthreads();
```

now adjacent threads can do adjacent accesses into `data`, eg `data[threadIdx.x]`!

Shared memory as cache

```
// For CSR row multiplication example, adapted from P&H fig. B.8.5

__shared__ float cache[blocksize];
unsigned int block_begin = blockIdx.x * blockDim.x;
unsigned int block_end = block_begin + blockDim.x;
unsigned int row = block_begin + threadIdx.x;
if(row<num_rows) cache[threadIdx.x] = x[row];
__syncthreads();
...
// when reading each x_j
if (j >= block_begin && j < block_end)
    x_j = cache[j-block_begin];
else
    x_j = x[j];
```

Not guaranteed locality (as we were in matrix multiplication), but increases performance as long as multiplying row *i* accesses values of x near *x[i]*

Shared memory banks

Shared memory is banked (32 banks for 32 threads/warp; successive words in successive banks)

really fast as long as no bank conflicts (have to do an extra round of accesses for every conflict – can significantly slow down warp)

Which code is better for working with data of length `n = 2 * blocksize` when `i = threadIdx.x`?

```
A[i * 2] = A[i * 2] + B[i * 2] // 0, 2, 4, 6… 2n - 2
A[i * 2 + 1] = A[i * 2 + 1] + B[i * 2 + 1] // 1, 3, 5, 7… 2n - 1
```

vs

```
A[i] = A[i] + B[i] // 0, 1, 2, 3, … n - 1
A[n + i] = A[n + i] + B[n + 1] // n, n + 1, n + 2, … 2n - 1
```

What's wrong with our CSR mult?

```
void csrMult(int n, int* Rp, int* C, float* V, float* x, float* y) {
    int r = blockIdx.x * blockDim.x + threadIdx.x;
    if (r < n) {
        int rBeg = Rp[r];
        int rSize = Rp[r + 1] - Rp[r];
        float sum = 0
        for (int i = 0; i < rSize; i++) {
            sum += V[rBeg + i] * x[C[rBeg + i]];
        }
        y[r] = sum;
    }
}
```

Solution: pad and transpose

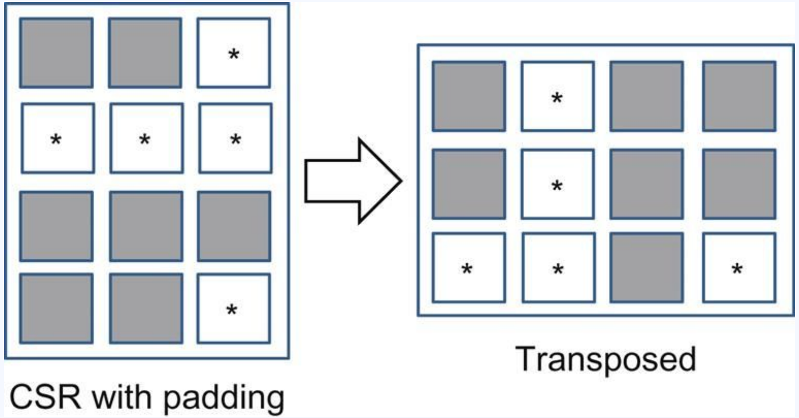
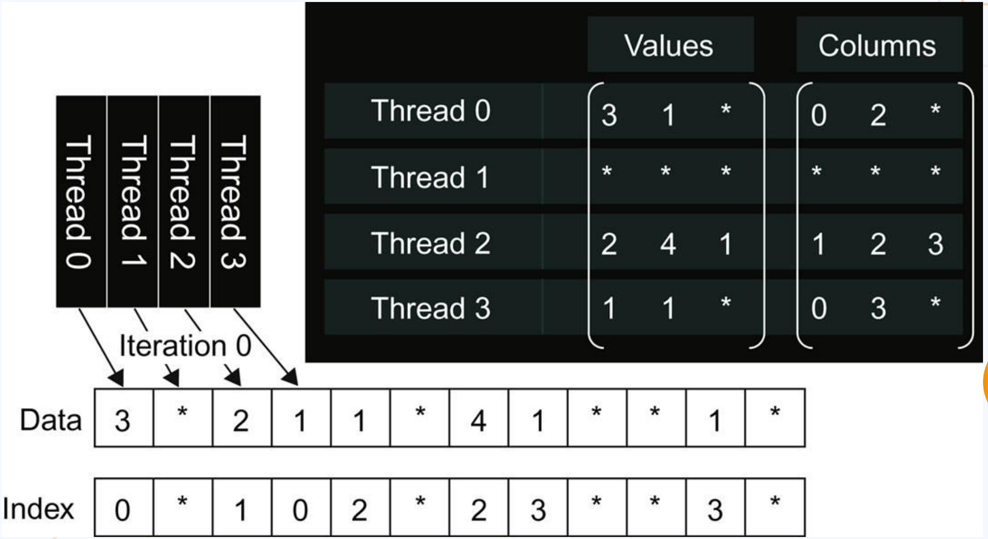


Image source: Kirk, David B., and W. Hwu Wen-Mei. *Programming massively parallel processors: a hands-on approach*. Morgan kaufmann, 2016., figs 10.8 and 10.9
Brown library access



Padding: allows for avoiding control flow divergence
Transpose: allows for coalescing
In general will run faster, despite extraneous multiplies by 0

? ? ?

__syncthreads() synchronizes all threads within a block.
How can threads in different blocks safely communicate with each other?

CPU/GPU communication

Must copy between CPU and GPU memory before/after launching kernel

Full code

```
int main() // on host
{
    // Allocate input vectors h_A and h_B in host memory
    float* h_A = (float*)malloc(size);
    // Initialize input vectors
    ...
    // Allocate vectors in device memory
    float* d_A;
    cudaMalloc(&d_A, size);
    // Copy vectors from host memory to device memory
    cudaMemcpy(d_A, h_A, size, cudaMemcpyHostToDevice);
```

Unified memory

GPU can access CPU's memory

Traditionally used with integrated graphics (GPU on same SoC as CPU)

CUDA example

Pros/cons?

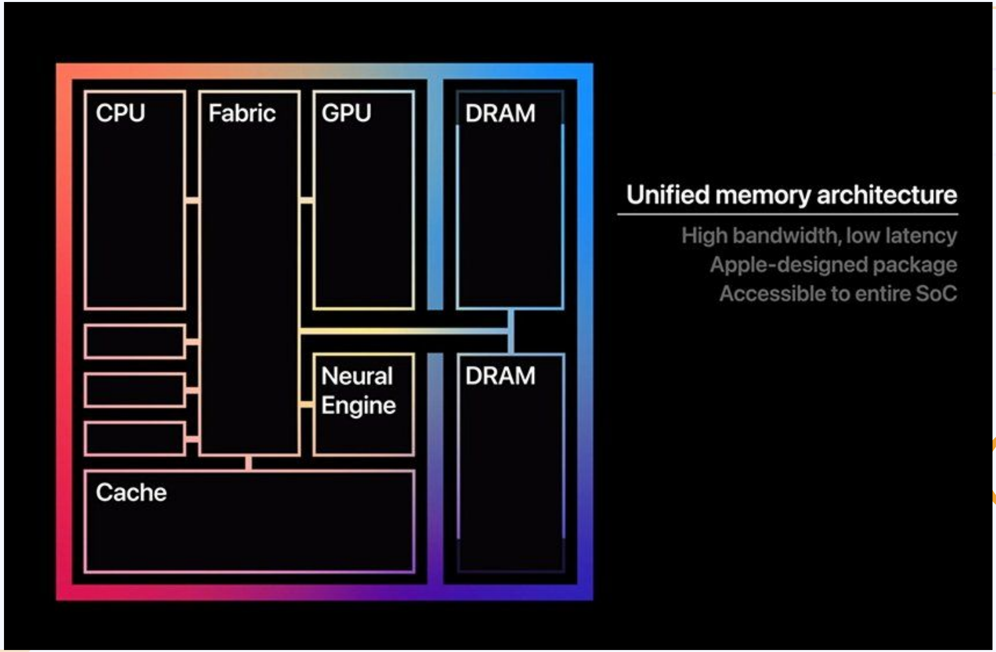


image source: Apple, inc

HW/SW interface

Vector processors and SIMD multimedia modifies the CPU ISA to support DLP. BUT:

- Supports only modest levels of parallelism (for SIMD extensions)
- Requires changes to ISA
- Requires compiler that can effectively vectorize code (or a skilled programmer)

SPMD model/GPUs: Allows programmer to write a kernel, which the hardware schedules on many threads on GPU. BUT:

- Proper performance requires proper understanding of architecture (branching/control divergence, memory access, synchronization)
- Requires interaction of CPU/GPU (might be a pro or a con)