SITCON'21

SHARED MEMORY - DYNAMIC

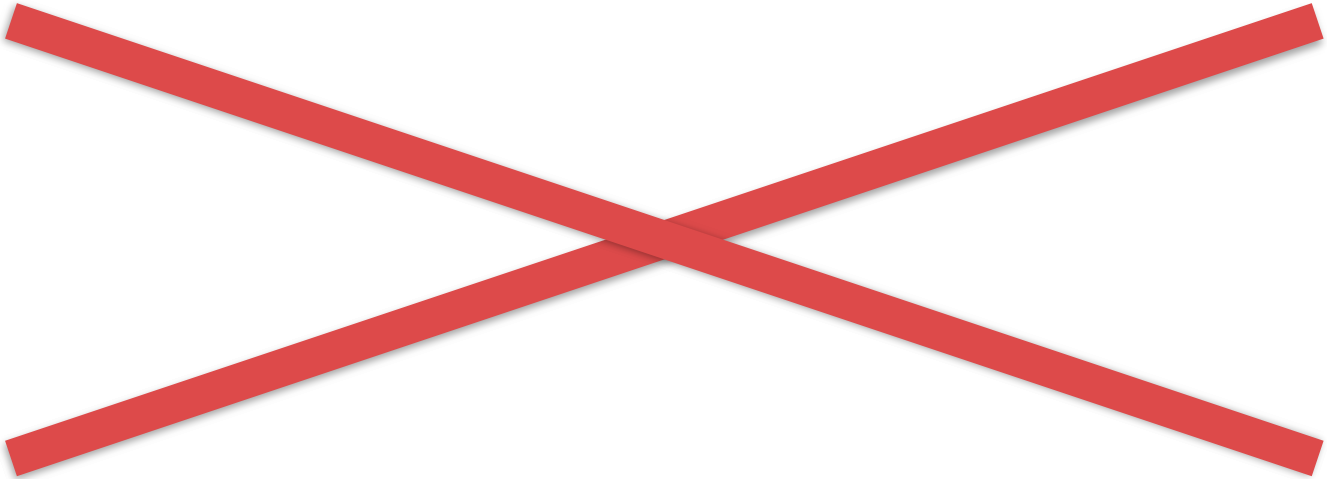
- `extern __shared__ int s[];`
- `func<<< block, thread, SM_size >>>();`

```
#define LEN 1000

__global__ void gpu_func(int *arr, int sz) {
    extern __shared__ int s[];
    int id = threadId.x;
    int bs = blockDim.x;
    for (int i = 0; i < LEN / bs; i++) {
        s[i * bs + id] = arr[i * bs + id];
    }
    __syncthreads();
    ...
}

int main(int argc, char *argv[]) {
    gpu_func<<< 10, 100, sizeof(int)*LEN >>>(gpu_arr, LEN);
    return 0;
}
```

```
extern __shared__ int *s;
```



SHARED MEMORY - DYNAMIC

- `extern __shared__ int s[];`
- `func<<< block, thread, SM_size >>>();`

```
1   #define LEN 1000
2
3   __global__ void gpu_func(int *arr, int sz) {
4       extern __shared__ int s[];
5       int id = threadId.x;
6       int bs = blockDim.x;
7       for (int i = 0; i < LEN / bs; i++) {
8           s[i * bs + id] = arr[i * bs + id];
9       }
10      __syncthreads();
11      ...
12  }
13
14  int main(int argc, char *argv[]) {
15      gpu_func<<< 10, 100, sizeof(int)*LEN >>>(gpu_arr, LEN);
16      return 0;
17  }
```

~~`extern __shared__ int *s;`~~

SITCON'21

WRITING FAST GPU PROGRAM

連續記憶體讀取	L1, L2 cache
存取相同記憶體	Shared memory
增加GPU使用率	Large block size
減少溝通次數	Copy larger memory block
Warp divergence	unroll loops…