

GPU Computing

Lab2

Flip di un'immagine

Multithreading su CPU

Esempio: flip (V/H) di un'immagine



flip orizzontale



flip verticale



Formato BitMap (BMP)

- ✓ Il formato di file Windows bitmap **lossless** di dimensioni **$W \times H$**
- ✓ Le immagini bitmap possono avere una **profondità** di 1, 4, 8, 16, 24 o 32 **bit x pixel**
- ✓ Ogni riga ha una lunghezza **N** in byte **multipla** di **4**, dove **$N = (3 * W + 3) \& (~3)$**
- ✓ Le bitmap con 1, 4 e 8 bit contengono una **tavolozza** per la **conversione** dei (rispettivamente 2, 16 e 256) possibili indici numerici nei rispettivi colori
- ✓ Nelle immagini con profondità più alta il **colore non è indicizzato** bensì codificato direttamente nelle sue componenti cromatiche **RGB**.
- ✓ La versione 3 del formato (in assoluto la più comune) non supporta il canale alfa né i metadati
- ✓ I **dati raw** sono così strutturati: **54 byte** occupati **dall'header** (width, height, inizio-fine tavolozza, etc.) poi seguono **3 byte per ogni pixel** (spazio RGB) della matrice di pixel che forma l'immagine

Formato BitMap (BMP)

- ✓ BMP files are **not compressed**, so each pixel takes up 3 bytes and it is possible to determine the **exact size** of the BMP files according to this formula:

$$Hbytes = (Hpixels \times 3 + 3) \wedge (11\dots1100)_2$$

$$24b\ RGB\ BMP\ File\ Size = 54 + Vpixels \times Hbytes$$

- ✓ It must be **rounded up** to the **next integer** that is **divisible by 4** to ensure that the BMP image size is a multiple of 4 by adding 3 and wiping out the LSB two bits of the resulting number
- ✓ The **conversion** from **Hpixels** to **Hbytes** should be as straightforward as

$$Hbytes = 3 \times Hpixels$$

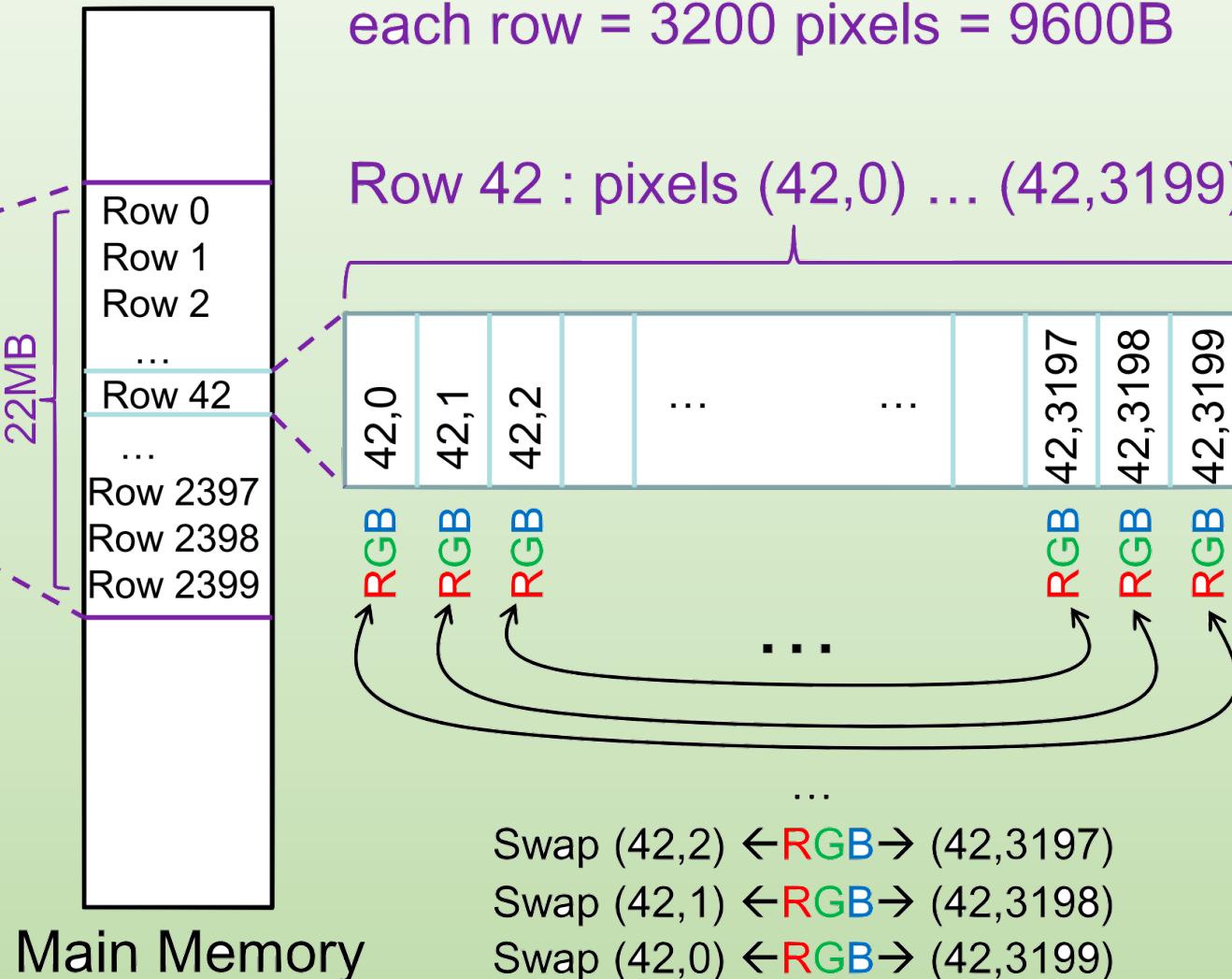
- ✓ Here are some **sample size** computations:
- ✓ A 24-bit **1024×1024** BMP need **3,145,782** bytes ($54 + 1024 \times 1024 \times 3$)
- ✓ A 24-bit **321×127** BMP need **122,482** bytes ($54 + (321 \times 3 + 1) \times 127$)

Rappresentazione

- ✓ **Struct** di definizione di un'immagine e funzioni lettura e scrittura

```
struct ImgProp {  
    int Hpixels;  
    int Vpixels;  
    unsigned char HeaderInfo[54];  
    unsigned long int Hbytes;  
};  
  
struct Pixel {  
    unsigned char R;  
    unsigned char G;  
    unsigned char B;  
};  
  
typedef unsigned char pel;      // pixel element  
  
pel** ReadBMP(char*);          // Load a BMP image  
void WriteBMP(pel**, char*);   // Store a BMP image
```

Schema di swap H



Pthreads library

POSIX-compliant operating system (Standard ANSI/IEEE POSIX 1003.1 - 1990)

Utilizzo

- ✓ Includere l'header della libreria `#include <pthread.h>`
- ✓ Compilare specificando la libreria: `gcc <opzioni> file_list -pthread`

API function

- ✓ `pthread_create()` permette di **creare** un nuovo **thread** e ne restituisce il suo **ID** (unico)
- ✓ `pthread_join()` mette in **attesa** il chiamante fino alla **terminazione** del **thread** (si deve sempre specificare il thread di cui si vuole attendere la terminazione, il suo uso è “obbligato” per garantire la terminazione coerente dei thread quando termina il processo che ha creati i thread)
- ✓ `pthread_attr()` consente di **inizializzare** gli **attributi** dei thread
- ✓ `pthread_attr_setdetachstate()` **assegna/modifica** gli **attributi** appena inizializzati ai thread

Inizializzazione dei thread

- ✓ `pthread_attr_init()` e `pthread_attr_setdetachstate()` dicono al SO che si è pronti a lanciare un gruppo di thread (es. l'attributo `detachstate` di un thread specifica se è joinable o detached – joinable lo è per default)

```
 . . .
#define MAXTHREADS 128
int NumThreads;                                // Total number of threads working in parallel
int ThParam[MAXTHREADS];                        // Thread parameters ...
pthread_t ThHandle[MAXTHREADS];                // Thread handles
pthread_attr_t ThAttr;                          // Pthread attributes
void (*FlipFunc)(pel** img);                   // Function pointer to flip the image
void* (*MTFlipFunc)(void *arg);                // Function pointer to flip the image, multi-th version
. . .
if (NumThreads > 1) {
    pthread_attr_init(&ThAttr);
    pthread_attr_setdetachstate(&ThAttr, PTHREAD_CREATE_JOINABLE);
```

Creazione e lancio di thread

- ✓ **pthread_create()** inizializza e attiva l'esecuzione all'atto della creazione stessa
 - Il terzo argomento **MTFlipFunc** dice al thread di eseguire la **funzione**
 - Il quarto argomento **ThParam[i]** è argomento della funzione (in genere fornisce il **'puntatore'** ai **dati** su cui applicare la funzione eseguita dal thread)

```
for (i = 0; i < NumThreads; i++) {  
  
    ThParam[i] = i;  
  
    ThErr = pthread_create(&ThHandle[i], &ThAttr, MTFlipFunc, (void *)&ThParam[i]);  
  
    if (ThErr != 0) {  
  
        printf("\nThread Creation Error %d. Exiting abruptly... \n", ThErr);  
  
        exit(EXIT_FAILURE);  
  
    }  
}
```

Esecuzione dei task (con thread)

- ✓ IL `main()` crea i thread e **assegna** un unico `tid` a ognuno a **runtime** (`ThHandle[i]` nel codice sotto) e invoca una **funzione** per ogni thread (usando il puntatore `MTFlipFunc` con parametro `ThParam[i]`)

```
 . . .
ThParam[i] = i;
ThErr = pthread_create(&ThHandle[i], &ThAttr, MTFlipFunc, (void *)&ThParam[i]);
```

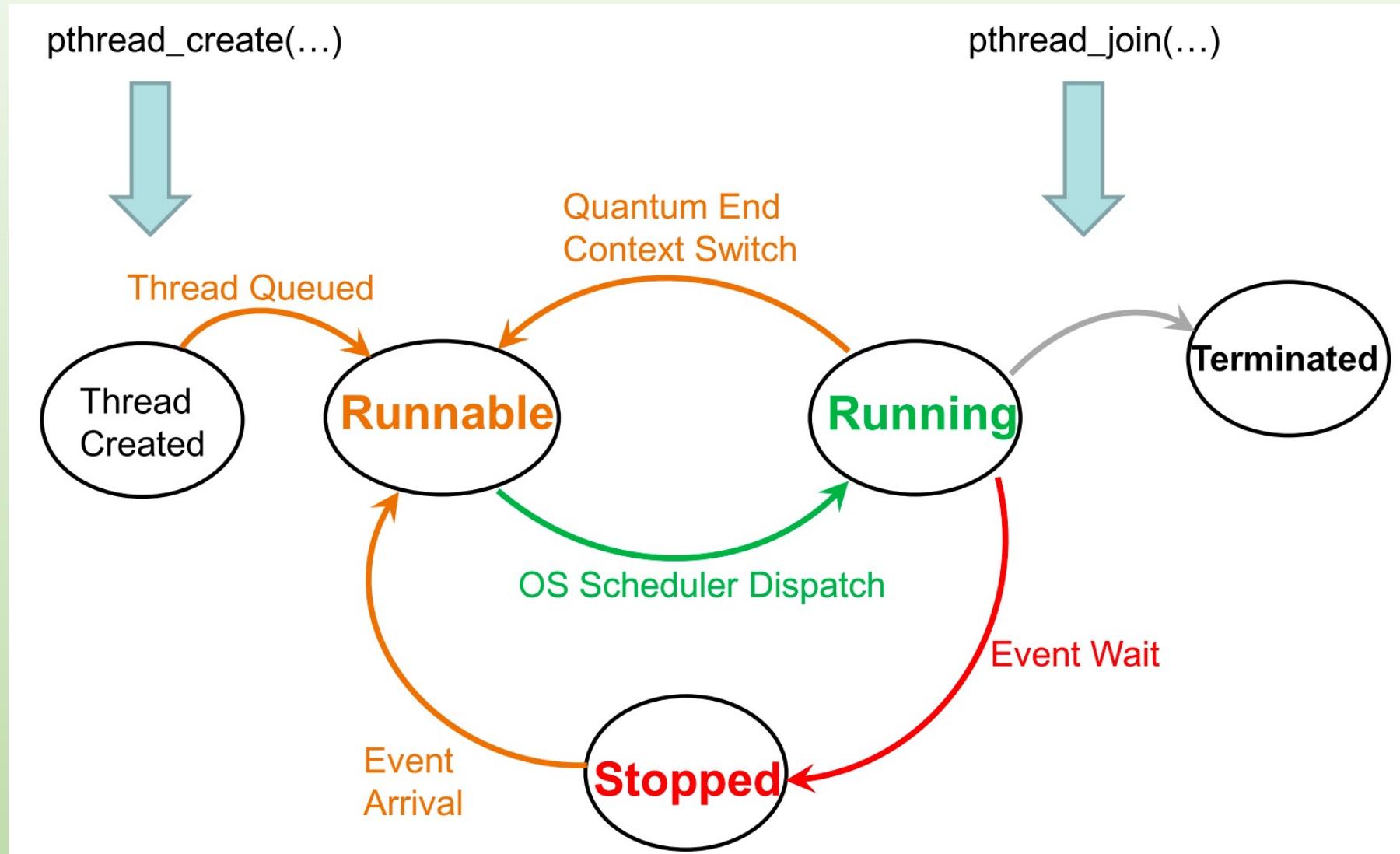
- ✓ Il `main()` deve anche **comunicare** al SO quali thread sono stati creati
- ✓ Il SO deve **decidere** se **creare** il thread (per es. se ci sono le risorse necessarie)
- ✓ Se un thread viene creato allora il SO gli **assegna** un handle in `ThHandle[i]`
- ✓ Alla fine il `main()` deve attendere che ogni thread termini – sincronizzazione `join` (dice all'OS di deallocate le risorse)

Join di thread

- ✓ Forma elementare di **sincronizzazione**: un thread si mette in **attesa** di un **altro thread**
- ✓ il thread che effettua il join si **blocca** finché uno specifico **thread non termina**
- ✓ Primo parametro **ThHandle[i]** è l'**id** del **thread** su cui sincronizzare (wait)
- ✓ **pthread_attr_destroy()** significa: *destroys the given thread-attribute object*
- ✓ **pthread_join(x)** significa: *wait until thread with the handle number x is done*
- ✓ La funzione **restituisce 0** in caso di **successo**, un valore diverso da zero in caso contrario

```
. . .
pthread_attr_destroy(&ThAttr);
for (i = 0; i < NumThreads; i++) {
    pthread_join(ThHandle[i], NULL);
}
. . .
```

Ciclo di vita di un pthread



Flip verticale: divisione dei dati

- ✓ Se abbiamo un'immagine di dimensione 640 x 480 la versione sequenziale deve fare mirroring tra righe a partire da quelle in posizione estreme:

```
Row [0]: [0][0]↔[479][0], [0][1]↔[479][1] ... [0][639]↔[479][639]
Row [1]: [1][0]↔[478][0], [1][1]↔[478][1] ... [1][639]↔[478][639]
Row [2]: [2][0]↔[477][0], [2][1]↔[477][1] ... [2][639]↔[477][639]
Row [3]: [3][0]↔[476][0], [3][1]↔[476][1] ... [3][639]↔[476][639]
...
Row [239]: [239][0]↔[240][0], [239][1]↔[240][1] ... [239][639]↔[240][639]
```

- ✓ nella versione parallela devo ripartire il compito tra i vari thread... per es. se ho 4 threads, `tid = 0,1,2,3`

<code>tid = 0 : Pixels [0...159]</code>	<code>Hbytes [0...477]</code>
<code>tid = 1 : Pixels [160...319]</code>	<code>Hbytes [480...959]</code>
<code>tid = 2 : Pixels [320...479]</code>	<code>Hbytes [960...1439]</code>
<code>tid = 3 : Pixels [480...639]</code>	<code>Hbytes [1440...1919]</code>

```
void *MTFlipV(void* tid) {
    struct Pixel pix;                                //temp swap pixel
    int row, col;
    long ts = *((int *) tid);                        // My thread ID is stored here
    ts *= ip.Hbytes / NumThreads;                   // start index
    long te = ts + ip.Hbytes / NumThreads - 1; // end index
    for (col = ts; col <= te; col += 3) {
        ...
    }
}
```

Prestazioni

- ✓ Da che cosa dipendono le differenze?
- ✓ ... velocità della CPU, numero di core, numero di thread?... gerarchie di memorie?
- ✓ Per es. su una CPU 4C/8T cosa succede se due thread eseguono sullo stesso core o in core differenti?
- ✓ come cambiano le prestazioni quando si lanciano 9 o più thread su CPU 4C/8T?
- ✓ politiche di scheduling?
- ✓ gestione di eccezioni: che accade se un lancio di thread fallisce per effetto del SO?

Feature	CPU1	CPU2	CPU3	CPU4	CPU5	CPU6
Name	i5-4200M	i7-960	i7-4770K	i7-3820	i7-5930K	E5-2650
C/T	2C/4T	4C/8T	4C/8T	4C/8T	6C/12T	8C/16T
Speed:GHz	2.5-3.1	3.2-3.46	3.5-3.9	3.6-3.8	3.5-3.7	2.0-2.8
L1\$/C	64KB	64KB	64KB	64KB	64KB	64KB
L2\$/C	256KB	256KB	256KB	256KB	256KB	256KB
shared L3\$	3MB	8MB	8MB	10MB	15MB	20MB
Memory	8GB	12GB	32GB	32GB	64GB	16GB
BW:GB/s	25.6	25.6	25.6	51.2	68	51.2

#Threads	CPU1	CPU2	CPU3	CPU4	CPU5	CPU6	
Serial	V	109	131	159	117	181	185
2	V	93	70	50	58	104	95
3	V	78	46	33	43	75	64
4	V	78	67	49	59	54	49
5	V	93	55	40	52	35	57
6	V	78	51	35	55	35	48
8	V	78	52	37	53	26	37
9	V		47	34	52	25	49
10	V			40		23	45
12	V			35		28	38

ID di grid e block

Kernel: indici e dimensioni

Kernel function: uso delle vars builtin

Dimensionare blocchi e griglia...
Identificatore = variabile libera

La CPU non aspetta... prosegue

Runtime: distruzione contesto associato a processo

```
#include <stdio.h>

/*
 * Mostra DIMs e IDs di grid, block e thread
 */
__global__ void checkIndex(void) {
    printf("threadIdx:(%d, %d, %d) blockIdx:(%d, %d, %d) "
        "blockDim:(%d, %d, %d) gridDim:(%d, %d, %d)\n",
        threadIdx.x, threadIdx.y, threadIdx.z,
        blockIdx.x, blockIdx.y, blockIdx.z,
        blockDim.x, blockDim.y, blockDim.z,
        gridDim.x,gridDim.y,gridDim.z);
}

int main(int argc, char **argv) {
    // definisce grid e struttura dei blocchi
    dim3 block(4);
    dim3 grid(3);
    // controlla dim. dal lato host
    printf("grid.x %d grid.y %d grid.z %d\n", grid.x, grid.y, grid.z);
    printf("block.x %d block.y %d block.z %d\n", block.x, block.y, block.z);
    // controlla dim. dal lato device
    checkIndex<<<grid, block>>>();
    // reset device
    cudaDeviceReset();
    return(0);
}
```

Esecuzione

Compilazione

```
$ nvcc -arch=sm_20 grid1D.cu -o grid1D
```

Esecuzione con parametri **gridDim = 3** e **blockDim = 4**

```
$ ./ grid1D
grid.x = 3  grid.y = 1  grid.z = 1
block.x = 4  block.y = 1  block.z = 1
threadIdx:(0, 0, 0) blockIdx:(1, 0, 0) blockDim:(4, 1, 1) gridDim:(3, 1, 1)
threadIdx:(1, 0, 0) blockIdx:(1, 0, 0) blockDim:(4, 1, 1) gridDim:(3, 1, 1)
threadIdx:(2, 0, 0) blockIdx:(1, 0, 0) blockDim:(4, 1, 1) gridDim:(3, 1, 1)
threadIdx:(3, 0, 0) blockIdx:(1, 0, 0) blockDim:(4, 1, 1) gridDim:(3, 1, 1)
threadIdx:(0, 0, 0) blockIdx:(0, 0, 0) blockDim:(4, 1, 1) gridDim:(3, 1, 1)
threadIdx:(1, 0, 0) blockIdx:(0, 0, 0) blockDim:(4, 1, 1) gridDim:(3, 1, 1)
threadIdx:(2, 0, 0) blockIdx:(0, 0, 0) blockDim:(4, 1, 1) gridDim:(3, 1, 1)
threadIdx:(3, 0, 0) blockIdx:(0, 0, 0) blockDim:(4, 1, 1) gridDim:(3, 1, 1)
threadIdx:(0, 0, 0) blockIdx:(2, 0, 0) blockDim:(4, 1, 1) gridDim:(3, 1, 1)
threadIdx:(1, 0, 0) blockIdx:(2, 0, 0) blockDim:(4, 1, 1) gridDim:(3, 1, 1)
threadIdx:(2, 0, 0) blockIdx:(2, 0, 0) blockDim:(4, 1, 1) gridDim:(3, 1, 1)
threadIdx:(3, 0, 0) blockIdx:(2, 0, 0) blockDim:(4, 1, 1) gridDim:(3, 1, 1)
```

Completere il kernel grid2D

- ✓ “Filtrare” gli indici di un kernel CUDA basato su grid 2D che soddisfino il seguente requisito:
 - Attiva solo i thread con coordinate (sia nella componente x sia nella y) abbiamo per somma un multiplo di 5

```
#include <stdio.h>

/*
 * Show DIMs & IDs for grid, block and thread
 */
__global__ void grid2D(void) {
    // TODO
}

int main(int argc, char **argv) {
    // grid and block structure
    dim3 block(7,6);
    dim3 grid(2,2);

    // check for host
    printf("CHECK for host:\n");
    printf("grid.x = %d\t grid.y = %d\t grid.z = %d\n", grid.x, grid.y, grid.z);
    printf("block.x = %d\t block.y = %d\t block.z %d\n", block.x, block.y, block.z);

    // check for device
    printf("CHECK for device:\n");
    checkIndex<<<grid, block>>>();

    // reset device
    cudaDeviceReset();
    return (0);
}
```

Risultato grid2D

```
$ ./ grid2D
grid.x = 2 grid.y = 2 grid.z = 1
block.x = 8 block.y = 7 block.z 1
threadIdx:(1, 4, 0) blockIdx:(1, 0, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(6, 4, 0) blockIdx:(1, 0, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(0, 5, 0) blockIdx:(1, 0, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(5, 5, 0) blockIdx:(1, 0, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(4, 6, 0) blockIdx:(1, 0, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(1, 4, 0) blockIdx:(0, 1, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(6, 4, 0) blockIdx:(0, 1, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(0, 5, 0) blockIdx:(0, 1, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(5, 5, 0) blockIdx:(0, 1, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(4, 6, 0) blockIdx:(0, 1, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(1, 4, 0) blockIdx:(1, 1, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(6, 4, 0) blockIdx:(1, 1, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(0, 5, 0) blockIdx:(1, 1, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
threadIdx:(5, 5, 0) blockIdx:(1, 1, 0) blockDim:(8, 7, 1) gridDim:(2, 2, 1)
.
.
```

Flip di immagini con CUDA

Esercitazione: flipping di un'immagine

Sviluppare un programma CUDA C che effettua il flipping (V/H) di un'immagine

Osservazioni:

- ✓ Decidere che cosa deve fare l'host e che cosa il device
- ✓ la memoria dell'immagine è linearizzata 1D... tenerne conto nel disegno del kernel
- ✓ stabilire la dimensione di blocco di thread
- ✓ provare diverse configurazioni per aumentare le prestazioni
- ✓ misurare le prestazioni
- ✓ di seguito alcuni suggerimenti...

Flipping con CUDA: Soluzione 1D

parametro
libero

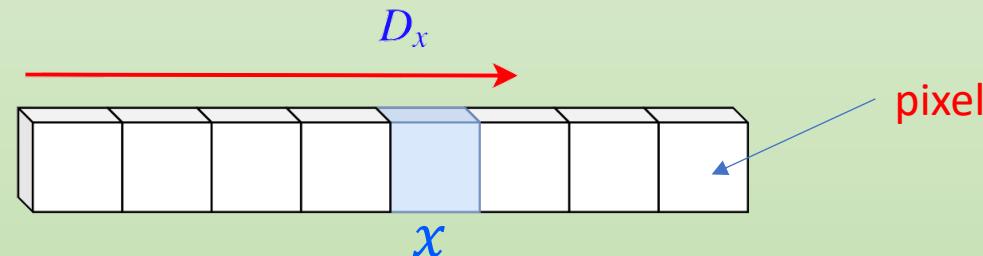
num blocchi
per riga

num blocchi
totali

```
// invoke kernels (define grid and block sizes)
dimBlock = 256;
rowBlock = (WIDTH + dimBlock - 1) / dimBlock;
dimGrid = HEIGHT * rowBlock;
...
Hflip<<<dimGrid, dimBlock>>>(GPUCopyImg, GPUImg, WIDTH);
...
Vflip<<<dimGrid, dimBlock>>>(GPUCopyImg, GPUImg, WIDTH, HEIGHT);
```

✓ OSS: La griglia indicizza I pixel... non la terna di i byte dei valori RGB!

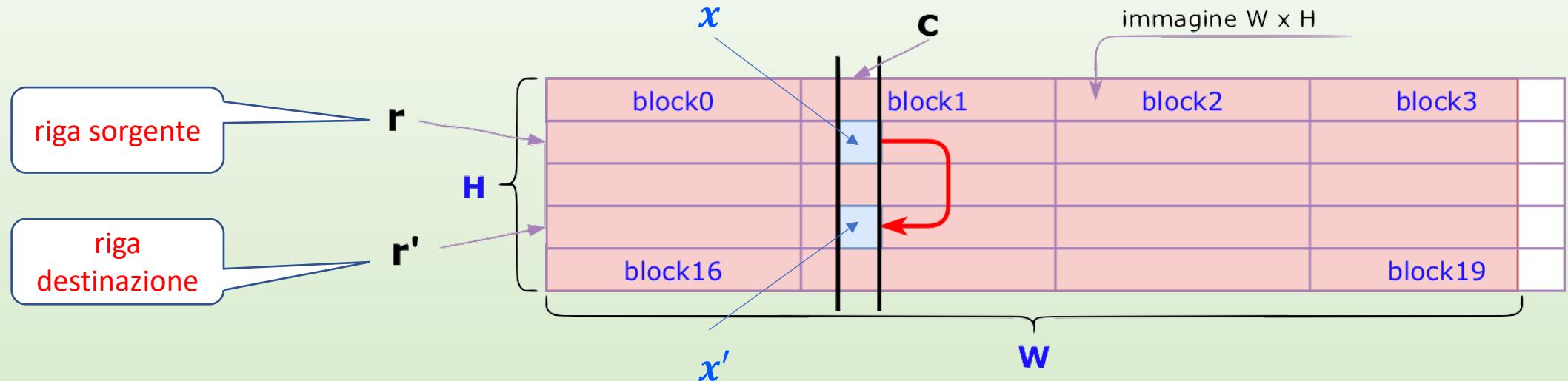
✓ Grid 1D e block 1D:



✓ Indice di pixel progressivo da 1 a $W \times H$
(W = 3200 H = 2400)

$$\begin{aligned}x &= blk_{Dim} * blk_{ID} + th_{ID} \\&= i * b + j\end{aligned}$$

Indici x flip verticale



✓ Indice del pixel sorgente:

$$x = \text{blk}_{\text{Dim}} * \text{blk}_{\text{ID}} + \text{th}_{\text{ID}}$$

$$x = |\text{blk}| * \text{blk}_i + \text{th}_j$$

$$x = b * i + j$$

✓ Indice del pixel da sostituire:

$$x' = r' * m + c'$$

num blocchi per riga:

riga sorgente:

colonna sorgente:

riga destinazione:

colonna destinazione:

$$m = (w + b - 1)/b = \left\lceil \frac{w}{b} \right\rceil$$

$$r = i/m$$

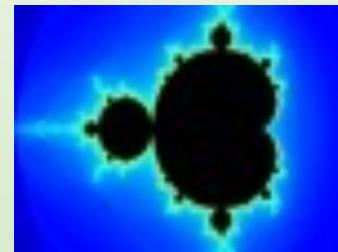
$$c = x - r * w$$

$$r' = h - 1 - r$$

$$c' = c$$

Flip verticale: colonne out of range

- ✓ Se si fissa la dimensione di blocco ($b = \text{blk}_{\text{Dim}} = |\text{blk}|$) la grid risultante ha un numero di colonne che potrebbe eccedere quello dell'immagine e quindi devono essere escluse



H=6000

W=8000



H=524

W=1024

$b = 64$

$$8000/b = 125$$

$$\text{num Blk} = b * 125 = 8000$$

$b = 128$

$$8000/b = 62.5$$

$$\text{num Blk} = b * 63 = 8064$$

$b = 256$

$$8000/b = 31.2$$

$$\text{num Blk} = b * 32 = 8192$$

eccedenti!!

$$1024/b = 16$$

$$\text{num Blk} = b * 16 = 1024$$

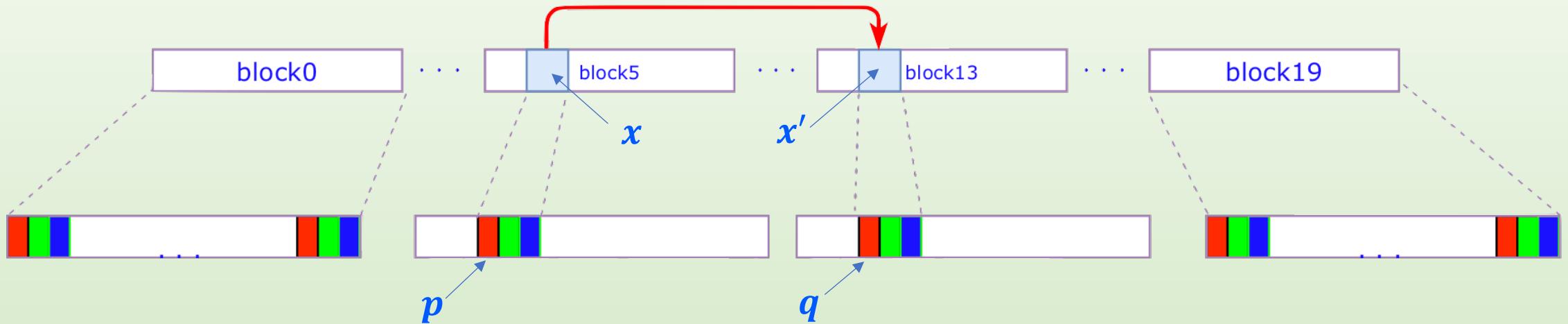
$$1024/b = 8$$

$$\text{num Blk} = b * 8 = 1024$$

$$1024/b = 4$$

$$\text{num Blk} = b * 4 = 1024$$

Accesso a byte in memoria lineare



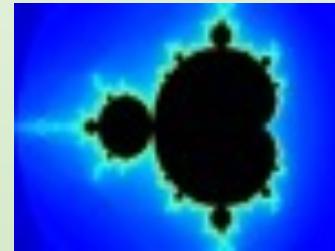
multiplo di 3
byte corrisp. a
terna RGB del
pixel src in
row r e col c

lo stesso per
pixel di dst in
row r' e col c

```
uint s = (w * 3 + 3) & (~3);           // num bytes x row (mult. 4)  
uint r1 = h - 1 - r;                   // dest. row (mirror)  
// ** byte granularity **  
uint p = s * r + 3*c;                 // src byte position of the pixel  
uint q = s * r1 + 3*c;                // dst byte position of the pixel  
// swap pixels RGB  
imgDst[q] = imgSrc[p];                // R  
imgDst[q + 1] = imgSrc[p + 1];        // G  
imgDst[q + 2] = imgSrc[p + 2];        // B
```

Tempi di computazione

- ✓ Tempi (in sec) di esecuzione di diverse **GPU** vs la **CPU** (2,9 GHz Intel Core i7 quad-core) e relativi **speedup** (rapporto tra tempi Dev/Hots)
- ✓ Immagini considerate:



Mandelbrot (W=8000, H=6000) = **144 MB**

Dog (W=1024, H=524) = **1.6 MB**

CPU/GPU	Dog	Mandelbrot	Speedup GPU vs CPU		
CPU	0.0026	0.3336		-	
Tesla M2090	0.00017		~15.2		-
Tesla K40	0.00011	0.0038	~23.6	-	~87.8
Tesla P100	0.000084	0.0014	~30.9	-	~238.3