Ejercicio 8

Observa que en el kernel de reducción que se presenta a continuación, para sumar N valores de un vector de números reales, la mitad de las hebras de cada bloque no hacen ningún trabajo después de participar en la carga de datos desde memoria global a un vector en memoria compartida (sdata). Modifica este kernel para eliminar esta ineficiencia y da los valores de los parámetros de configuración que permiten usar el kernel modificado para sumar N reales. ¿Habría algún costo extra en término de operaciones aritméticas necesitadas? ¿Tendría alguna limitación esta solución en términos de uso de recursos?

El codigo original es el siguiente:

```
__global__ void reduceSum(float *d_V,int N){
    extern __shared__ float sdata[];
    int tid = threadIdx.x;
    int i = blockIdx.x*blockDim.x + threadIdx.x;
    sdata[tid] = ((i < N) ? d_V[i] : 0.0f);
    __syncthreads();

    for (int s=blockDim.x/2; s>0; s>>=1) {
        if (tid < s) {
            sdata[tid] += sdata[tid + s];
        }
        __syncthreads();
    }
    if (tid == 0) d_V[blockIdx.x] = sdata[0];
}
```

La modificación que propongo es la siguiente:

```
__global__ void
reduceSum(float *d_V,int N)
{
    extern __shared__ float sdata[];

    int tid = threadIdx.x;
    int i = blockIdx.x*blockDim.x*2 + threadIdx.x;
    float suma = (i<n) ? d_V[i] : 0
    if ( i+blockDim.x < n)
        suma += d_V[i+blockDim.x]
    sdata[tid] = mySum;
    __syncthreads();

     for (int s=blockDim.x/2; s>0; s>>=1) {
        if (tid < s) {
            sdata[tid] = suma += sdata[tid + s];
        }
        __syncthreads();
```

```
    }
    if (tid == 0) d_V[blockIdx.x] = suma;
}
```

Con la modificación anterior lo que estamos realizando es una cosa muy simple. Ahorrar la mitad de las hebras. Ya que con $\frac{N}{2}$ somos capaces de traspasar todas los datos de *d_v* a memoria compartida. Ademas de traspasar los datos también realizamos la reducción a la mitad de N. Es decir realizamos la primera suma de los elementos que en el código anterior se realizaría en la primera interacción del for.