Trab 3 OCII

Thread CPU vs Thread CPU

Thread CPU (tradicional) : pesado, tem acesso a pilha própria e suporte para chamada recursiva, assim como área própria de memória
Thread GPU : leve, basicamente uma função

Thread,Warp,Bloco

Thread: uma linha de execução do programa
Bloco: um conjunto de threads, de tamanho a ser definido, todos os blocos devem ter o mesmo tamanho e um bloco enviado para a gpu só sairá após ser executado por completo
Warp: subdivisão do bloco contendo 32 threads consecutivos, executados sincronamente
Ordem de execução: Thread -> Warp -> Bloco

Blocos e SMX

A unidade SMX é responsável por quebrar os blocks em warps (grupos de 32 threads), a unidade SMX pode ter no máximo 64 warps ou 16 blocos alocados simultâneamente. A unidade pode executar 1 ou 2 instruções simultaneamente para warps distintos (contanto que não haja dependência de dados entre eles) devido aos 4 “warp controllers” existentes

Variáveis no código

Global: são os parâmetros da função global, tem como tempo de vida todo o escopo da gpu
Shared: variáveis declaradas com _shared_, são visíveis à todos os threads dentro do bloco, tem como tempo de vida todo o escopo do bloco
Local: variáveis declaradas dentro do thread, visíveis somente no escopo de instância , tem como tempo de vida o warp