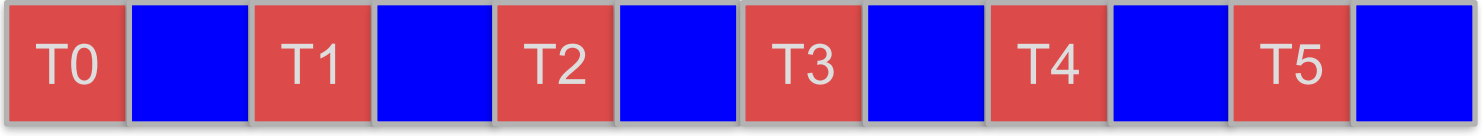
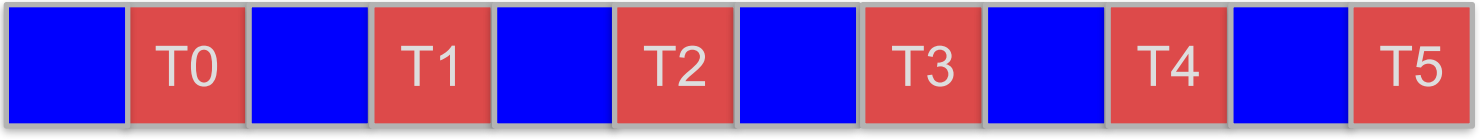
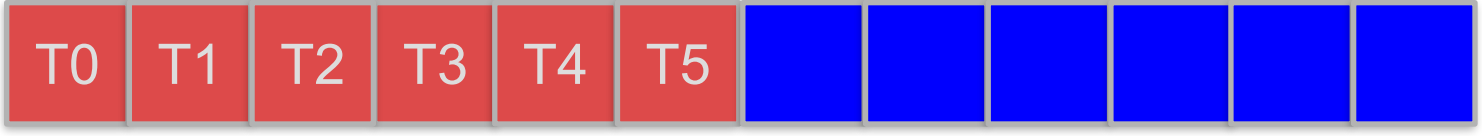
SITCON'21

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…


T0
T1
T2
T3
T4
T5

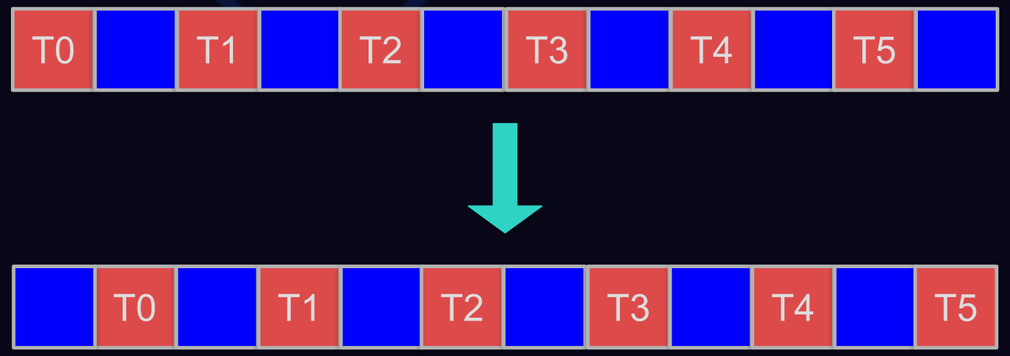

T0
T1
T2
T3
T4
T5

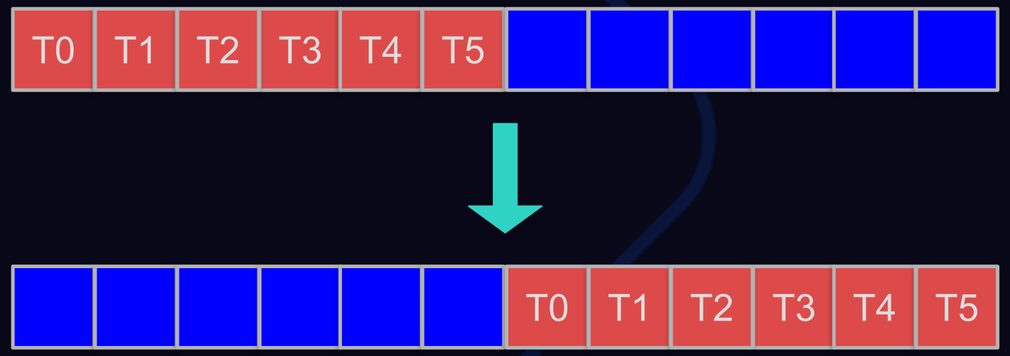

T0
T1
T2
T3
T4
T5


T0
T1
T2
T3
T4
T5

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…





DEMO

- All-Pairs Shortest Path
- 0/1 Knapsack

SITCON'21