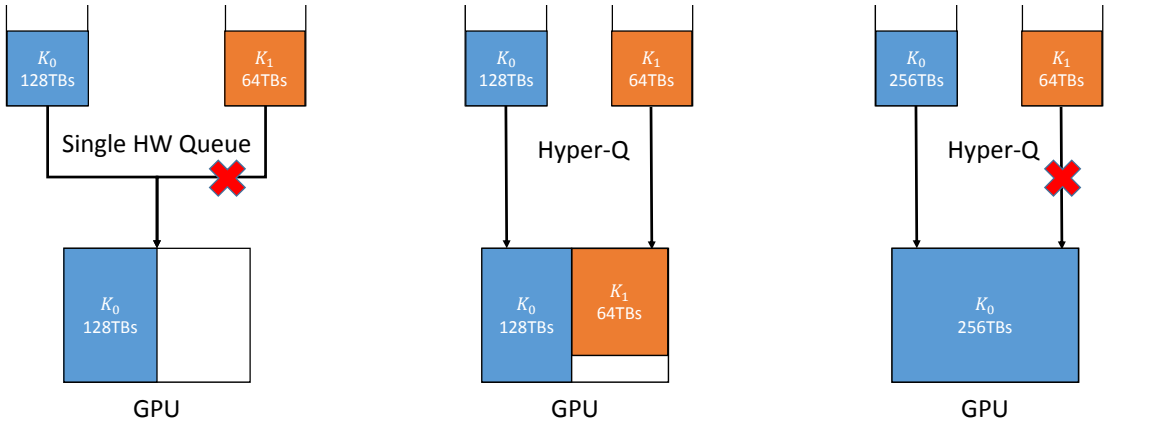
Simultaneous Multikernel GPU
(extracted from paper presentation in HPCA-15)

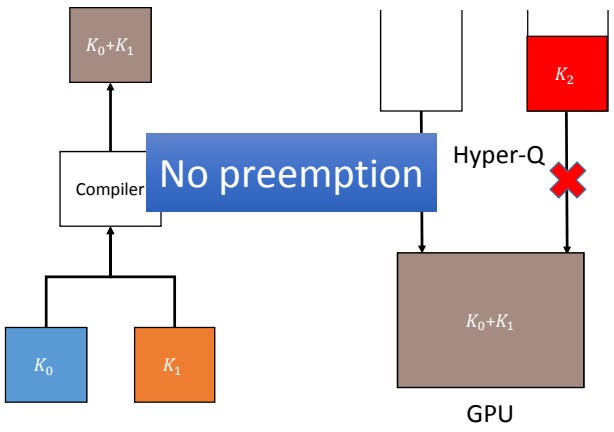
Multitasking GPU: Hyper-Q



HW queue limits parallelism

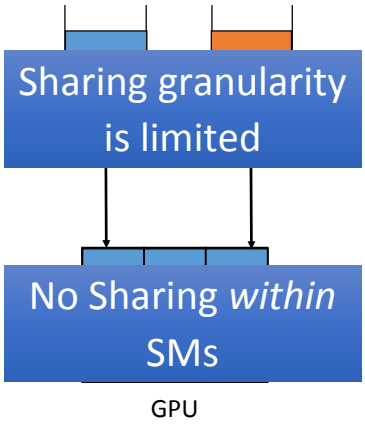
No control over sharing

Multitasking GPU: Previous SW/HW Designs



Elastic Kernel

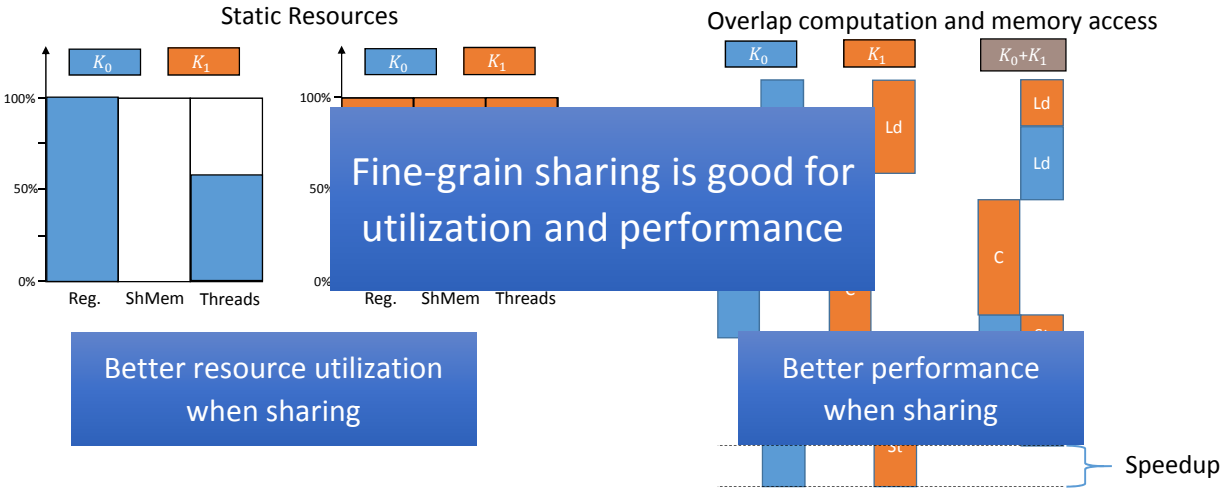
[ASPLOS '13]



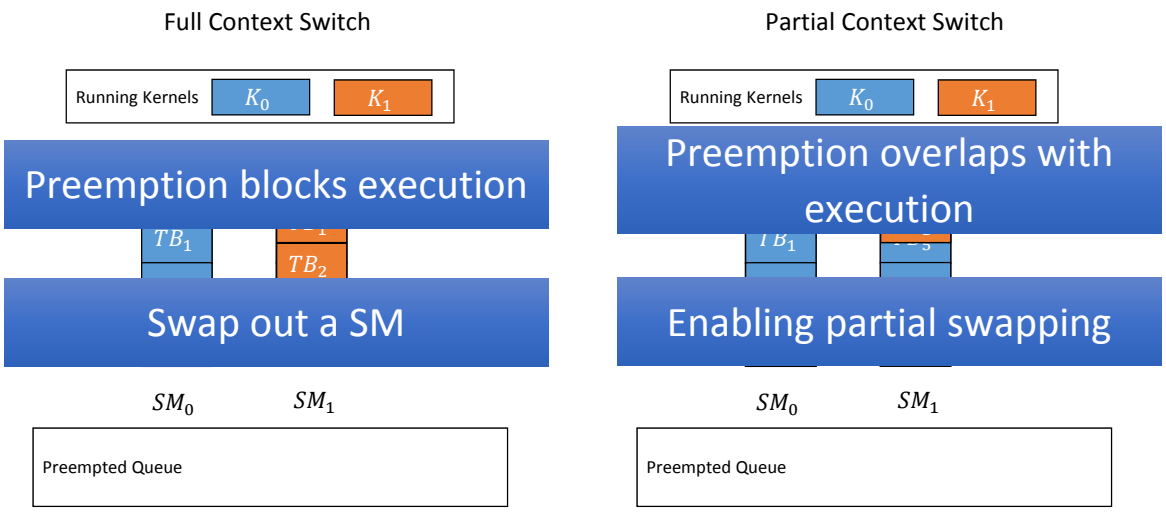
Spatial Partitioning

[ISCA '14]

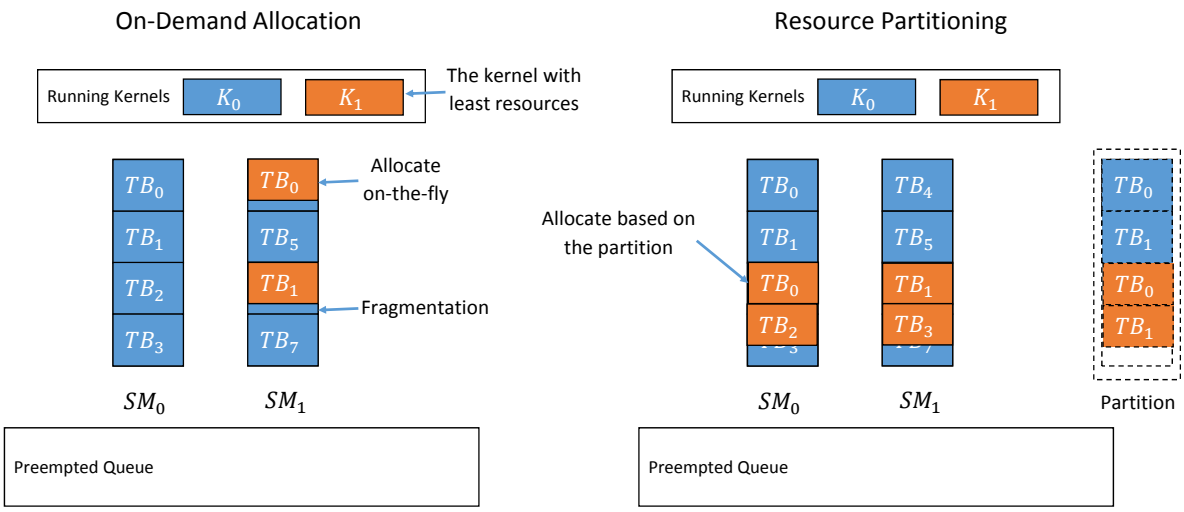
Motivation



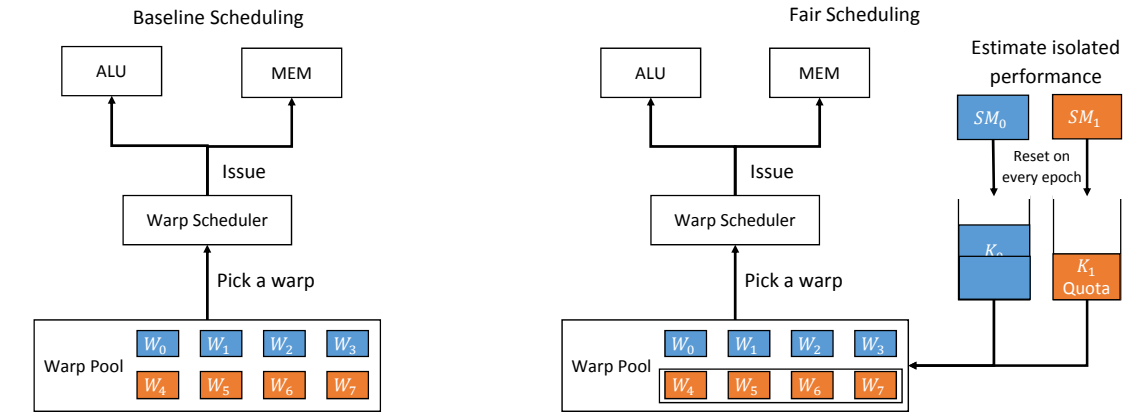
SMK: Partial Context Switch



SMK: Resource Allocation



SMK: Fair Warp Scheduling



Not aware of multikernel

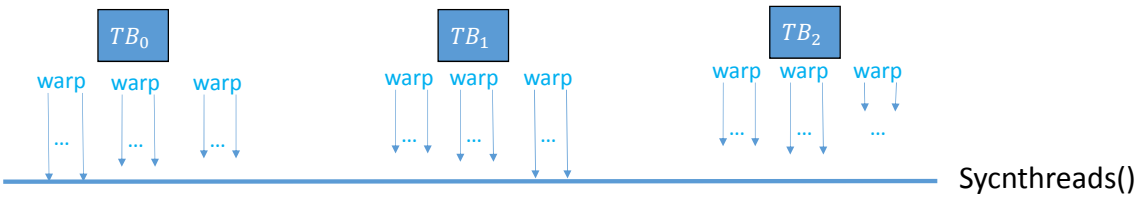
No performance fairness

Prevent over-issuing

Similar slowdown for kernels

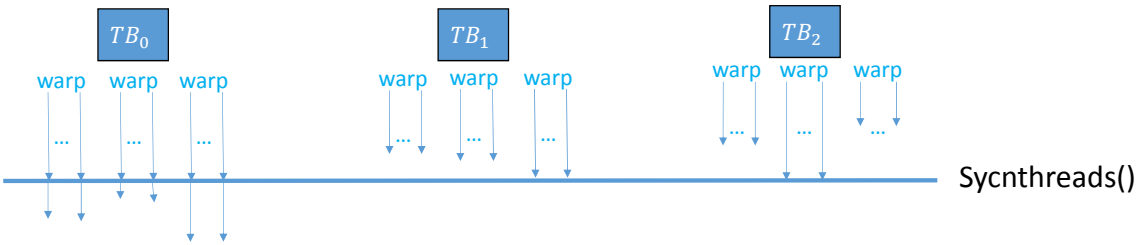
Barrier synchronization in GPU

- Barriers are only provided for threads within a TB.



Barrier synchronization in GPU

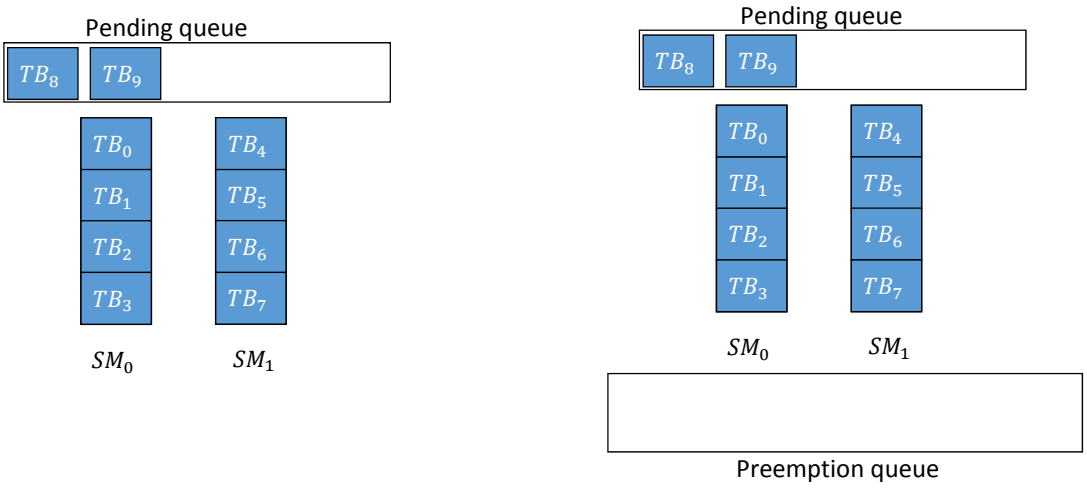
- Barriers are only provided for threads within a TB.



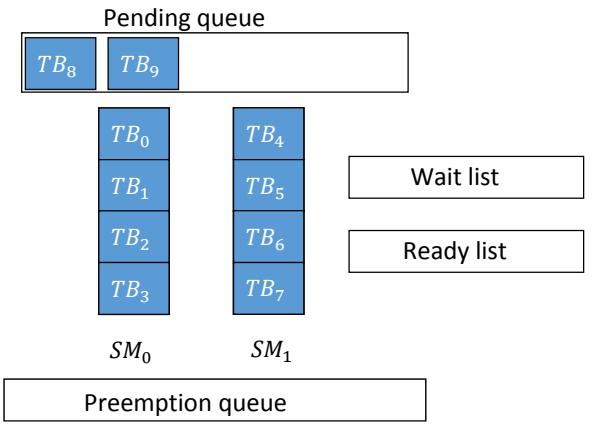
- No global barriers are provided in CUDA
- Need to use atomic global memory operations + busy-waiting to implement global barriers – but may deadlock.

Preemption allows barrier synchronization in GPU

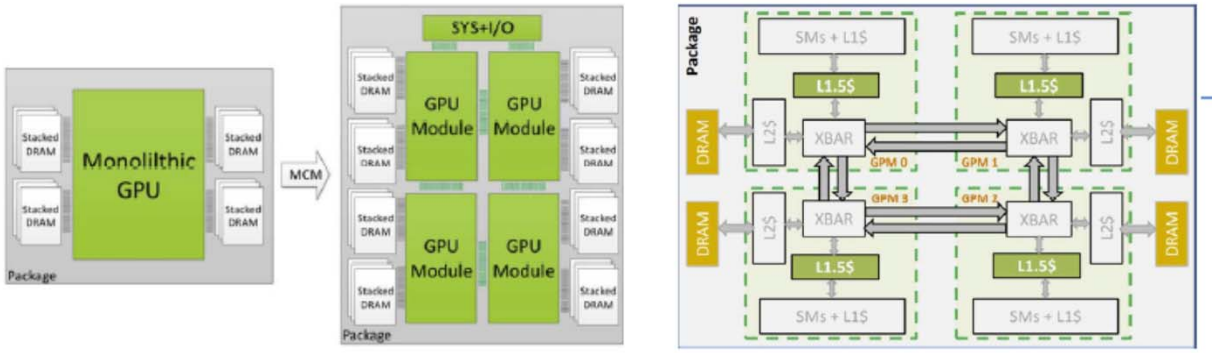
- No global (inter TB) barriers – will deadlock in the absence of preemption



Preemption allows wait/signal between TBs



Synchronizing the CPU with the GPU??



Nvidia GeForce card with multiple GPUs

https://www.tweaktown.com/news/58295/nvidia-shift-multiple-gpus-future-geforce-cards/index.html

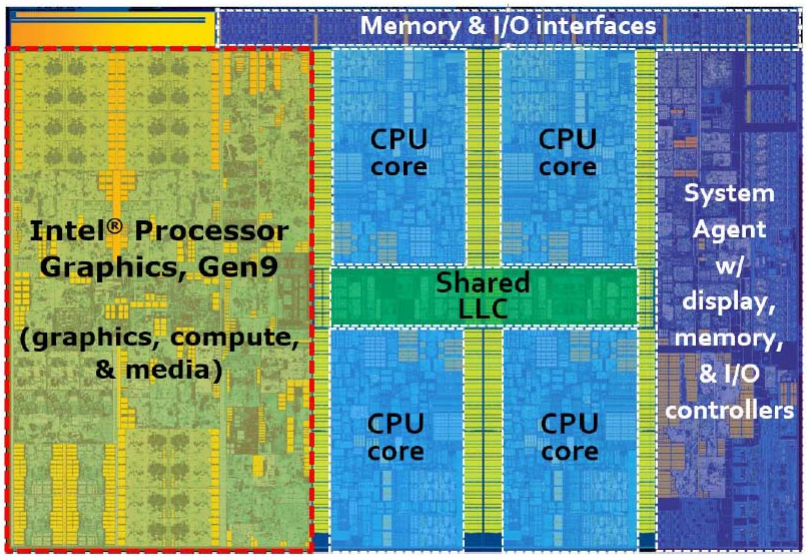
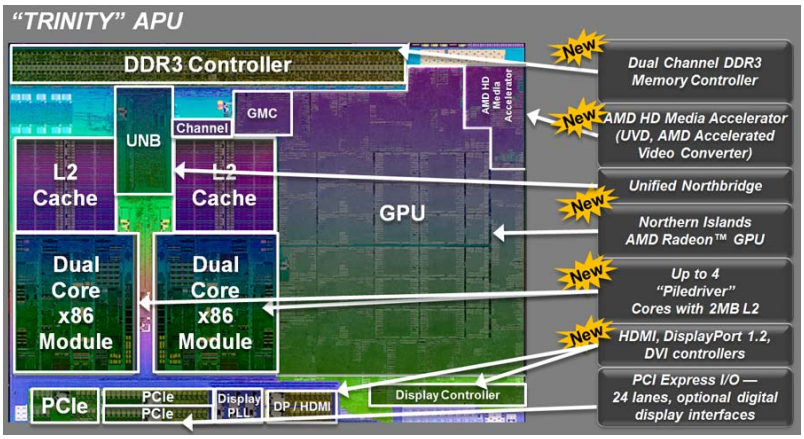


Figure 1: Architecture components layout for an Intel® Core™ i7 processor 6700K for desktop systems. This SoC contains 4 CPU cores, outlined in blue dashed boxes. Outlined in the red dashed box, is an Intel® HD Graphics 530. It is a one-slice instantiation of Intel processor graphics gen9 architecture.



Trinity's built-in AMD Radeon HD 7660D GPU includes 384 graphics cores (which AMD calls stream processors) running at 800MHz. It's worth noting that this number of stream cores gives the 7660D higher potential performance than AMD's entry-level discrete GPUs, such as the Radeon HD 7450, which only have 160 graphics cores. Given the emphasis on graphics, it's no surprise that the GPU takes up over half the Trinity die space.

http://www.pcworld.com/article/2010666/amd-announces-trinity-apus-superb-graphics-improved-cpu.html