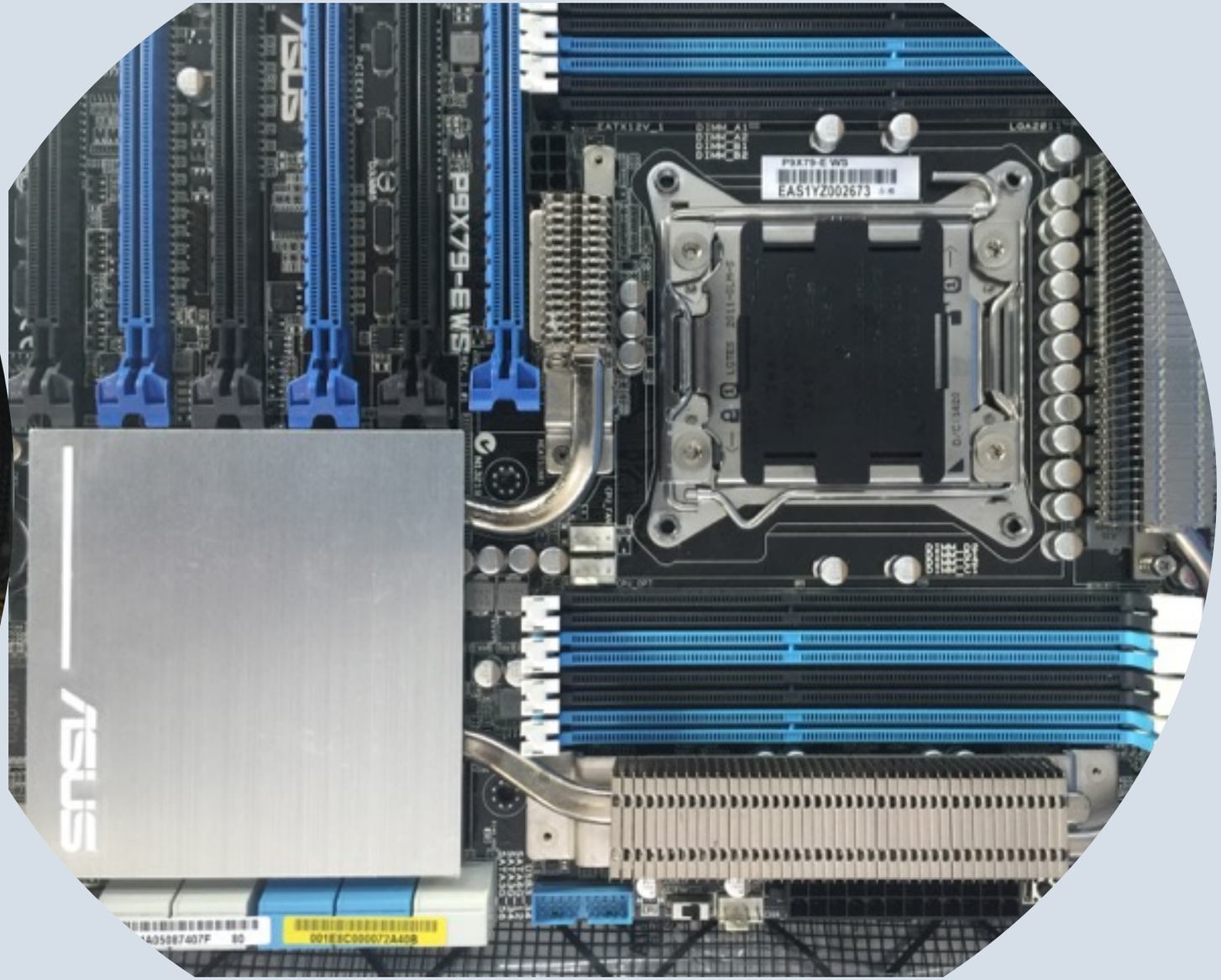
Lec 2 GPU Hardware Architecture

Dong Li, Tonghua Su
School of Software
Harbin Institute of Technology


GEFORCE GTX
NVIDIA TESLA K40
1200W


P9X79-E WS
EAS1YZ002673
ASUS

Outline

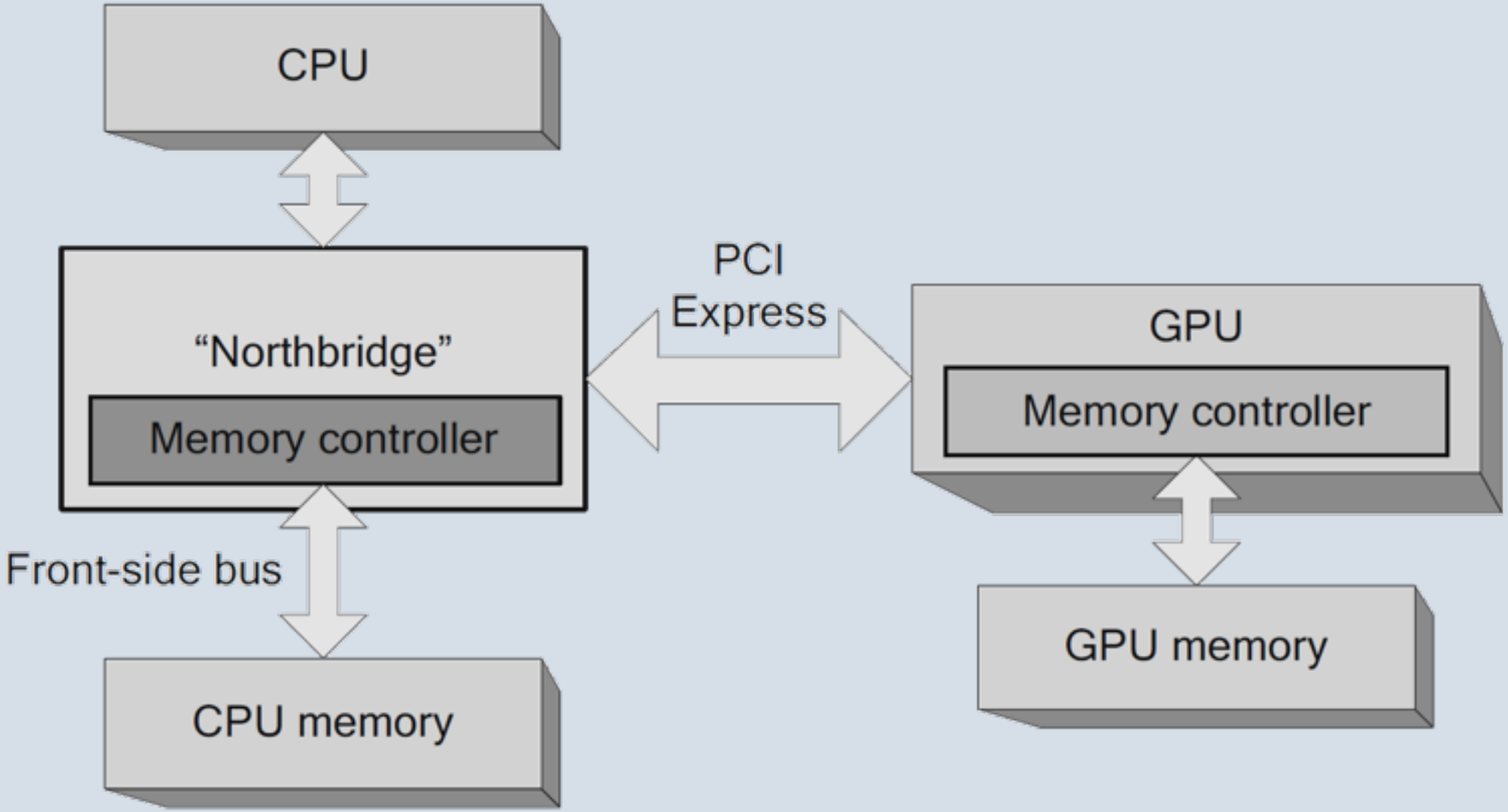
1. Linking Model
2. Kepler Architecture
3. Fermi Architecture

Outline

1. Linking Model
2. Kepler Architecture
3. Fermi Architecture

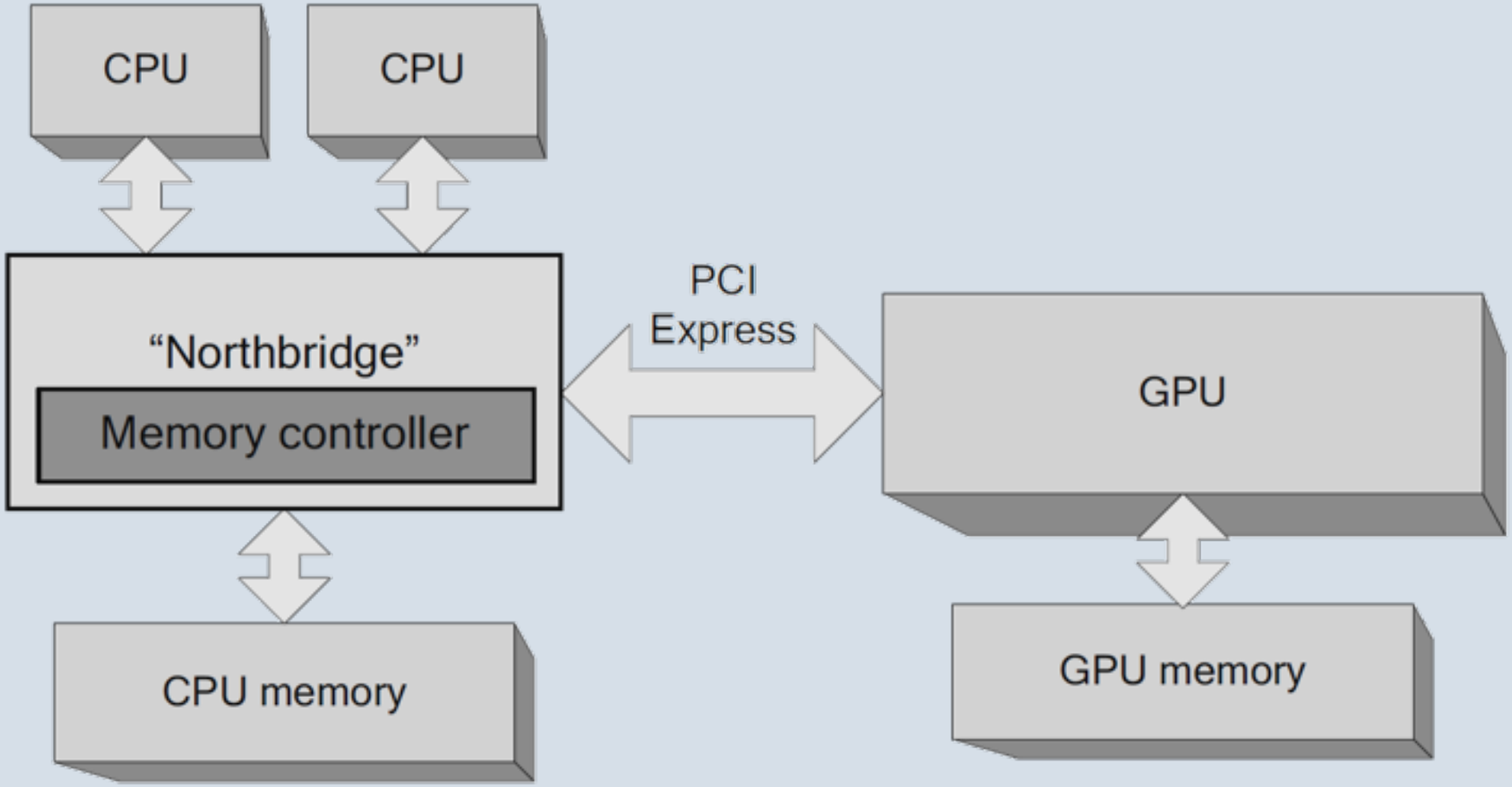
Linking Model

- **CPU/GPU architecture—northbridge**



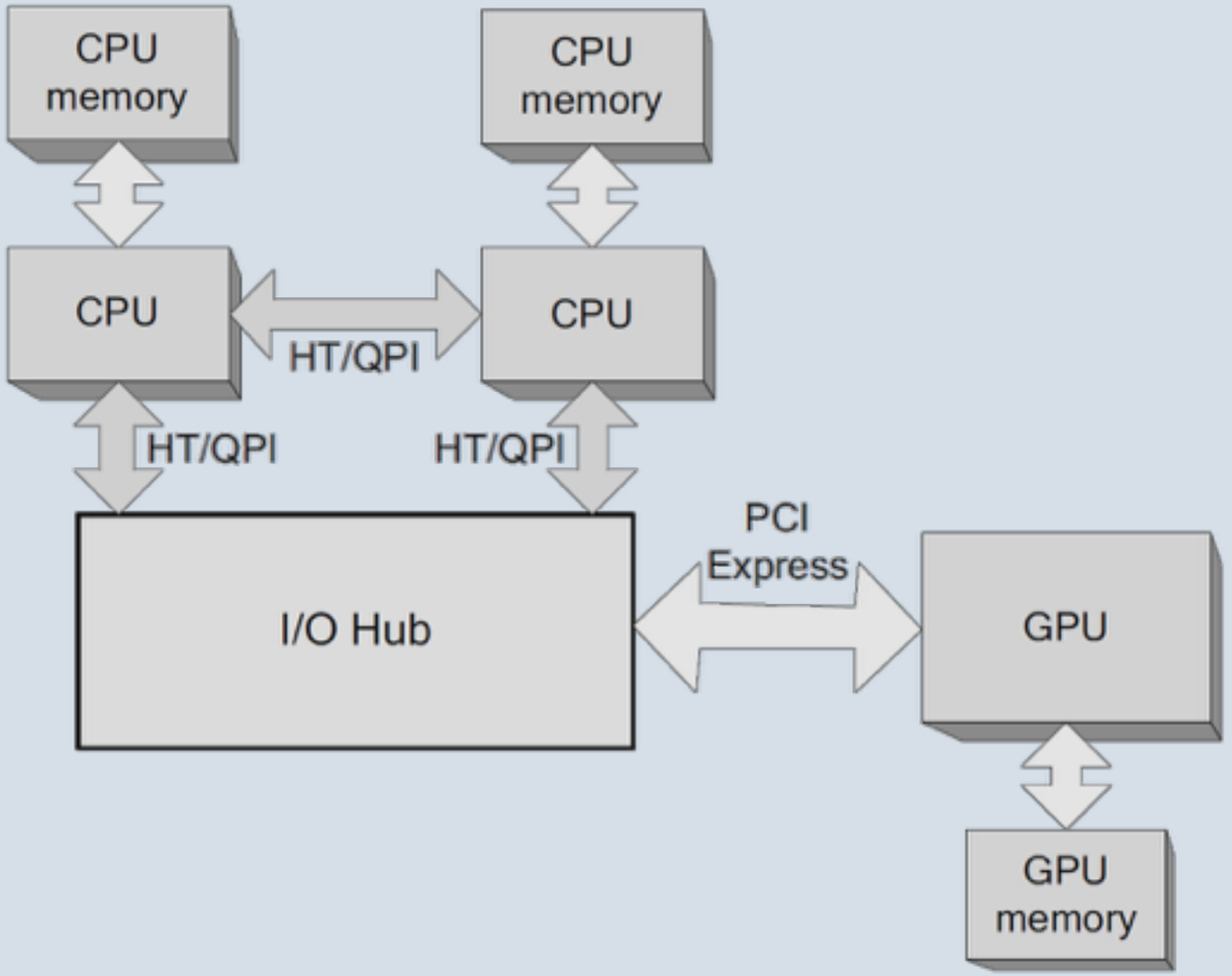
Linking Model

- **Multiple CPUs (SMP configuration)**



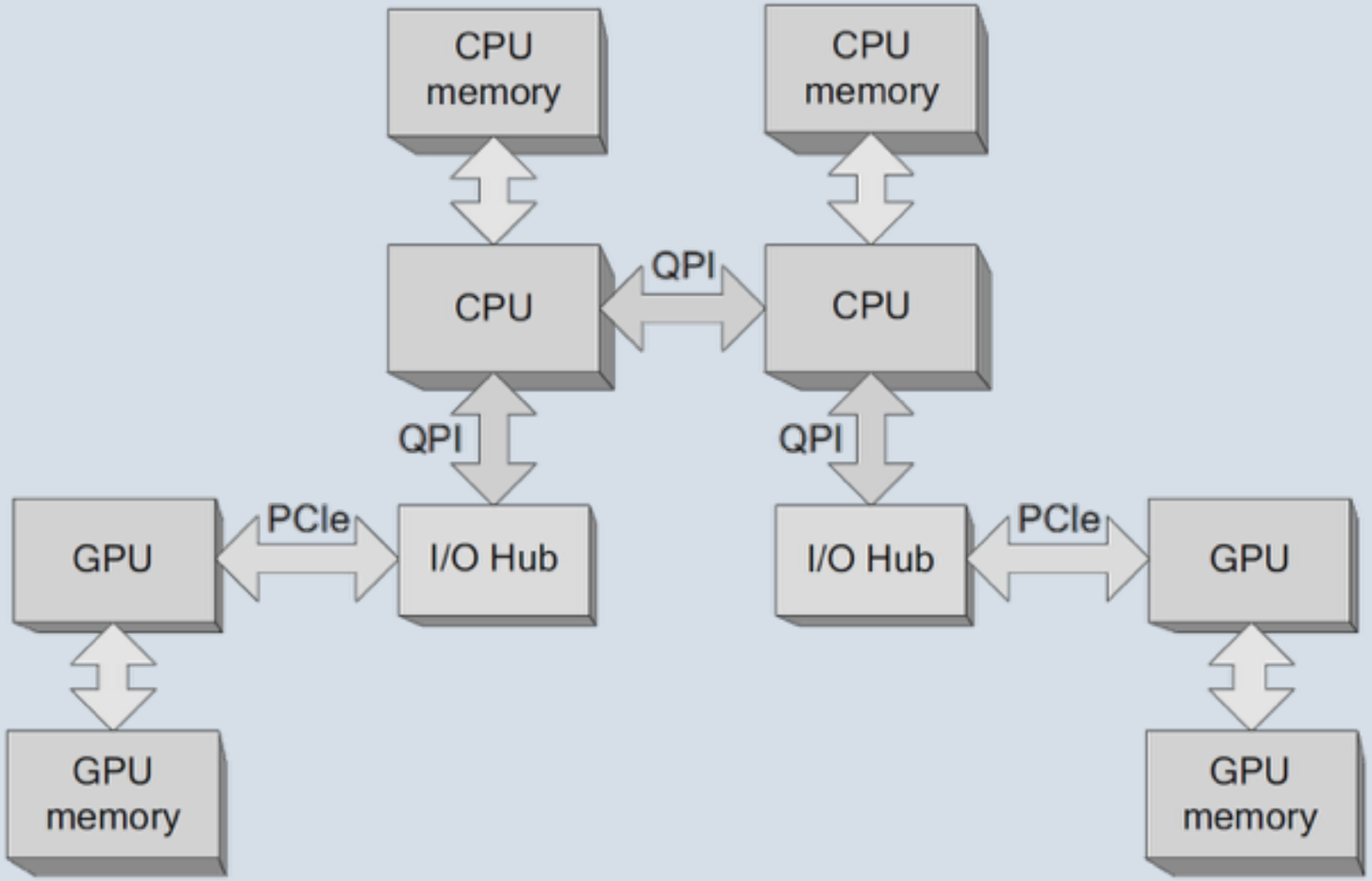
Linking Model

- **Multiple CPUs (NUMA)**



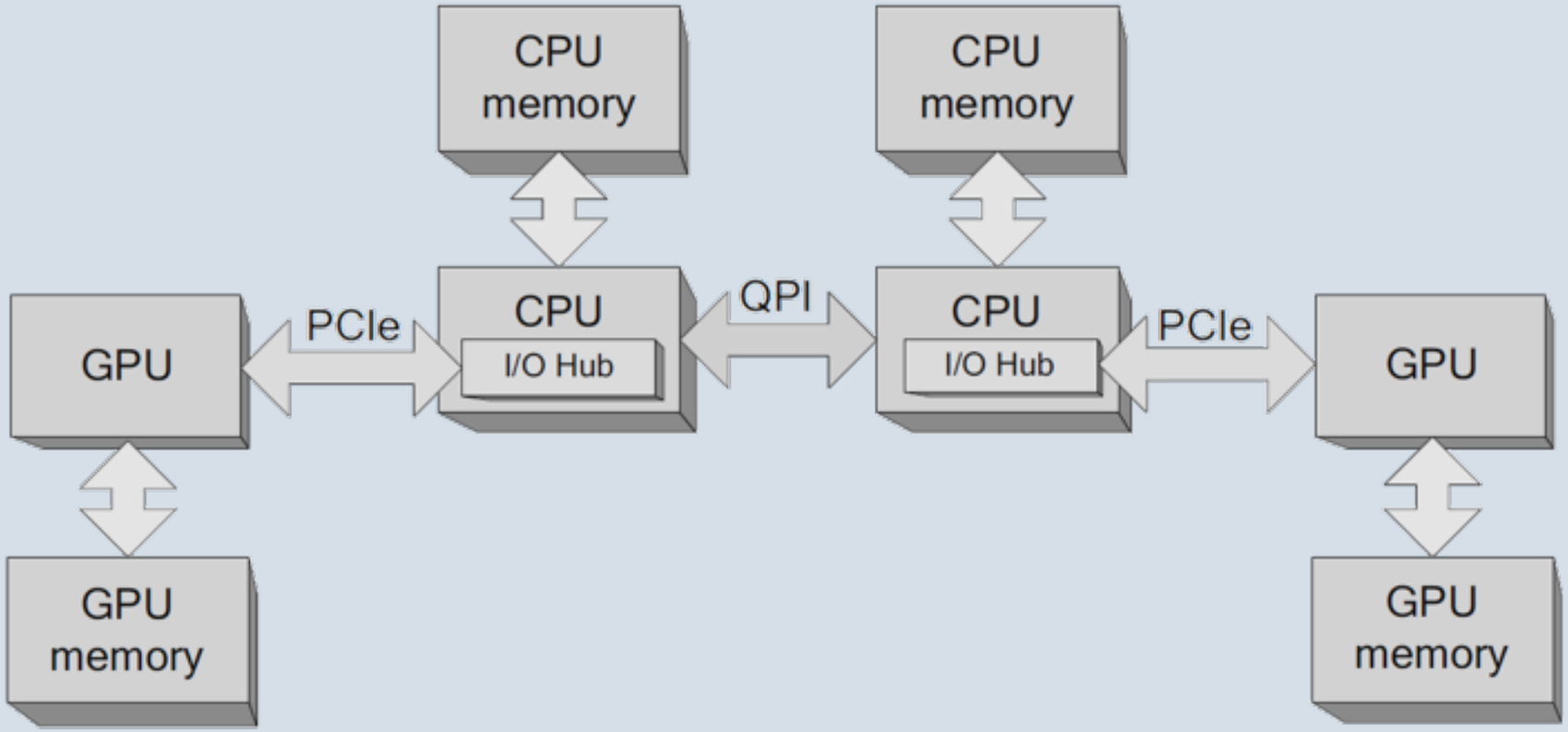
Linking Model

- **Multi-CPU (NUMA configuration), multiple buses**



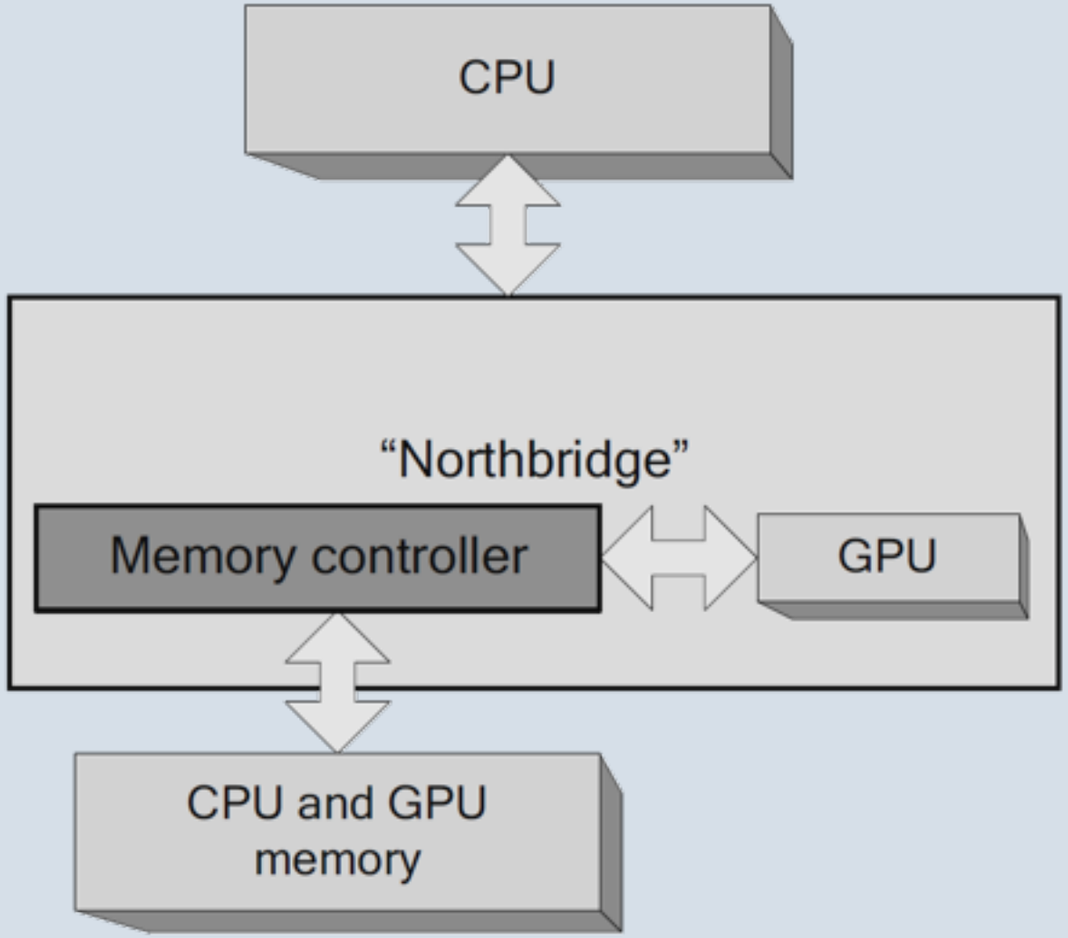
Linking Model

- **Multi-CPU with integrated PCI Express**



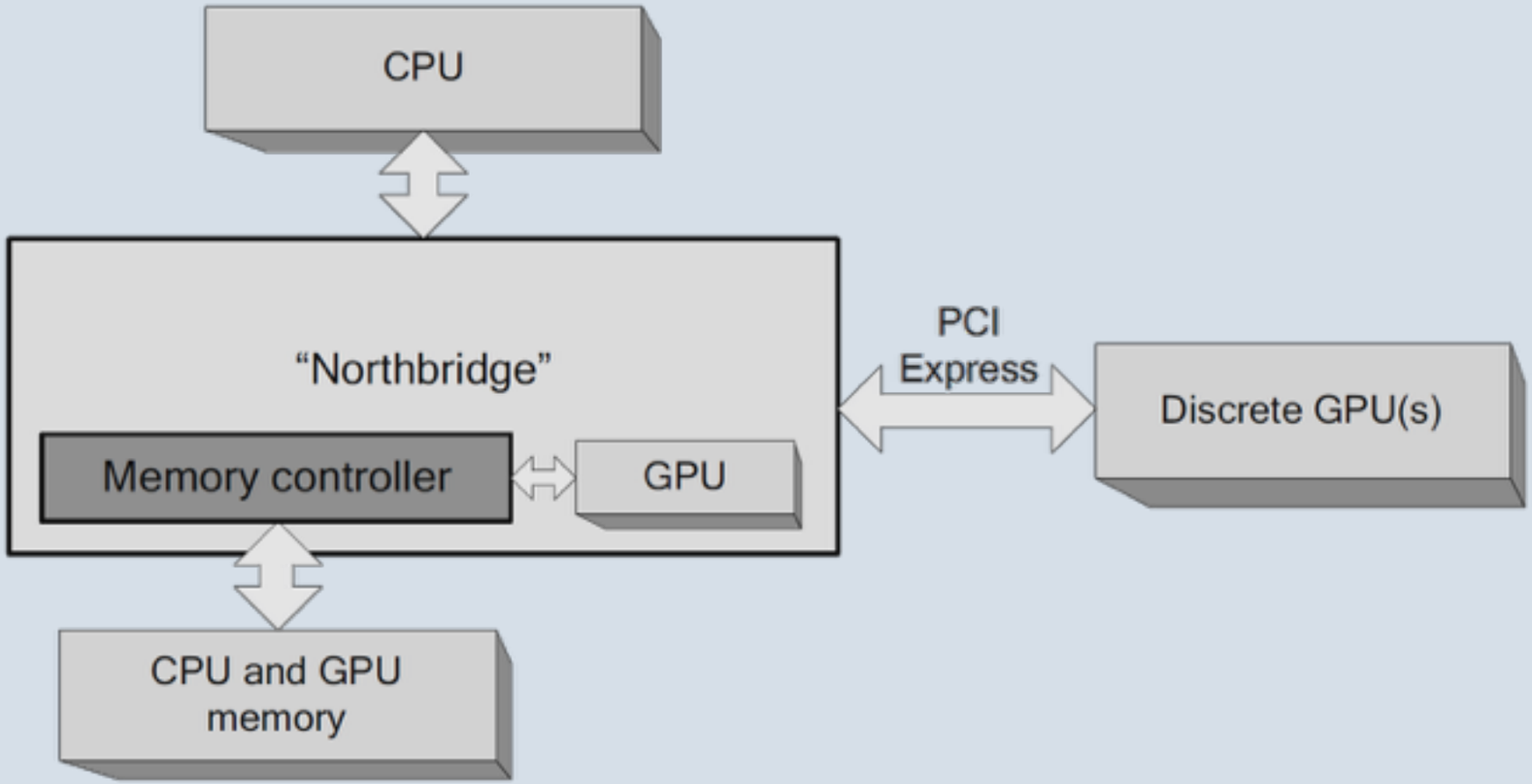
Linking Model

- **Integrated GPU**



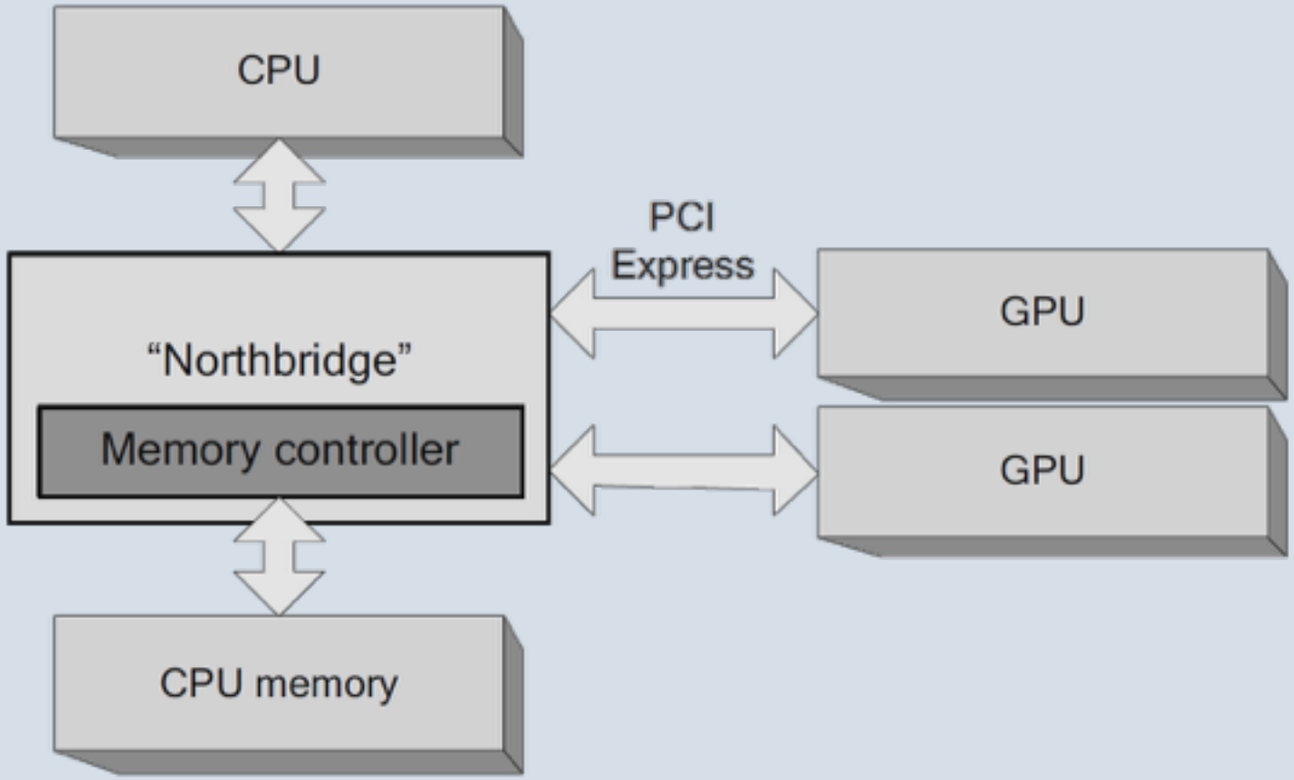
Linking Model

- **Integrated GPU with discrete GPU(s)**



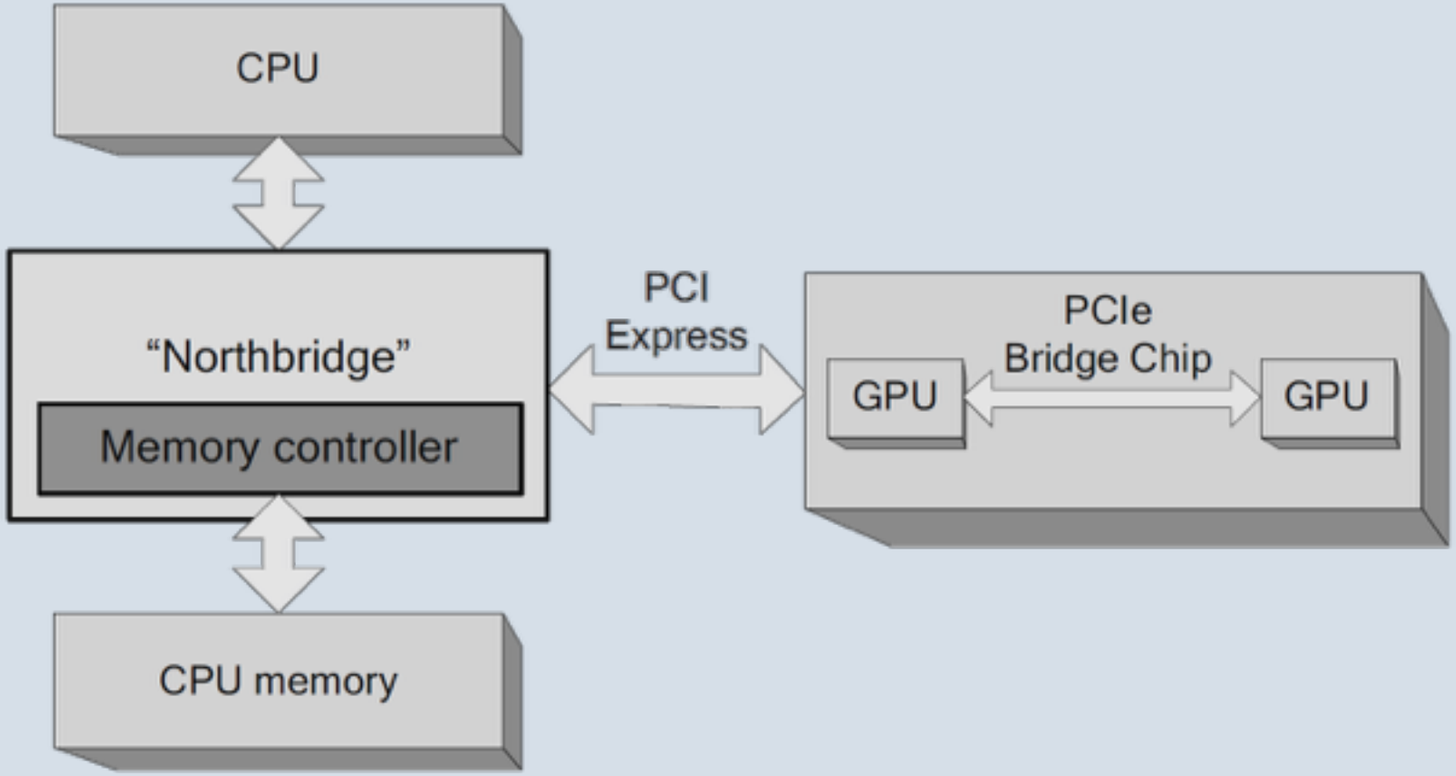
Linking Model

- GPUs in multiple slots



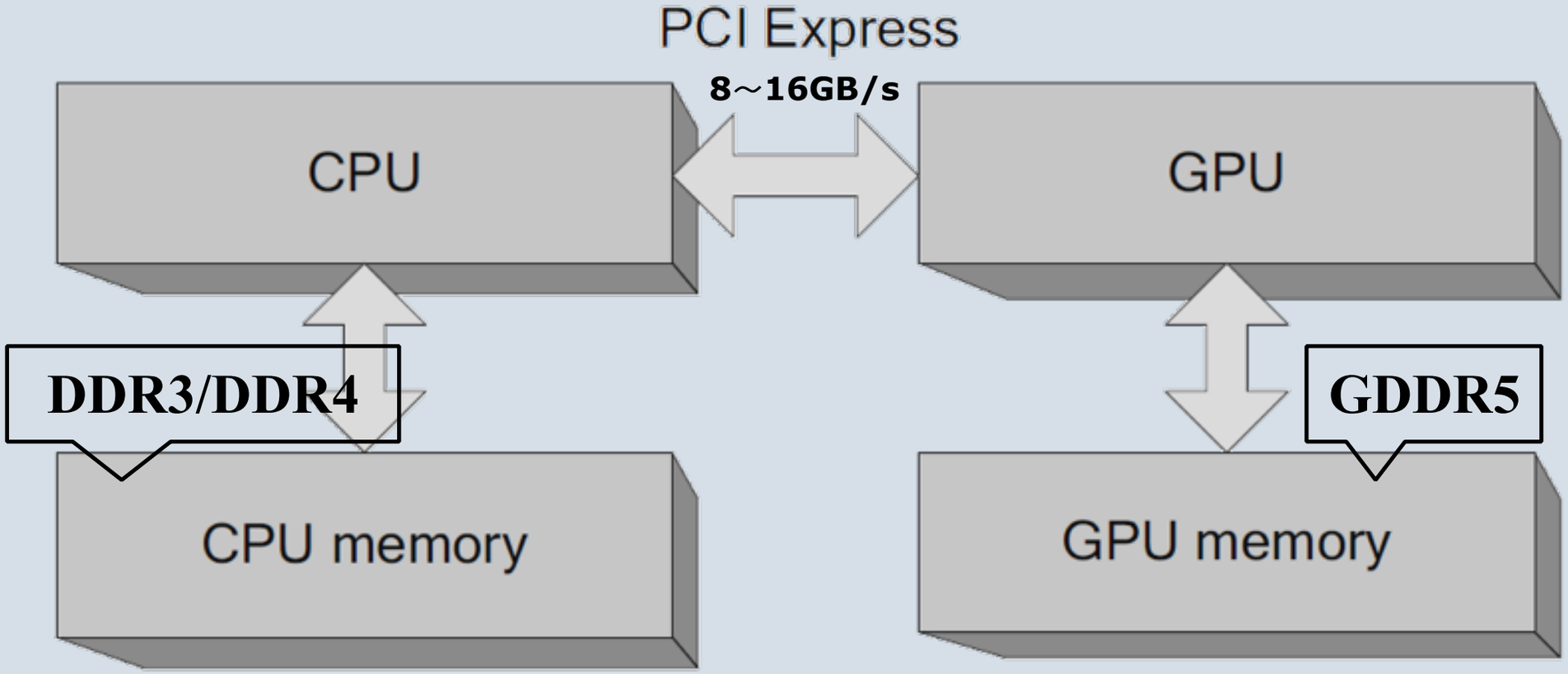
Linking Model

- **Multi-GPU board**



Linking Model

- **CPU/GPU architecture simplified**



Quiz

- 你的开发机器属于上述哪一种连接模型呢?
  - ✓ 请画出图示

Outline

1. Linking Model
2. Kepler Architecture
3. Fermi Architecture

Query Device

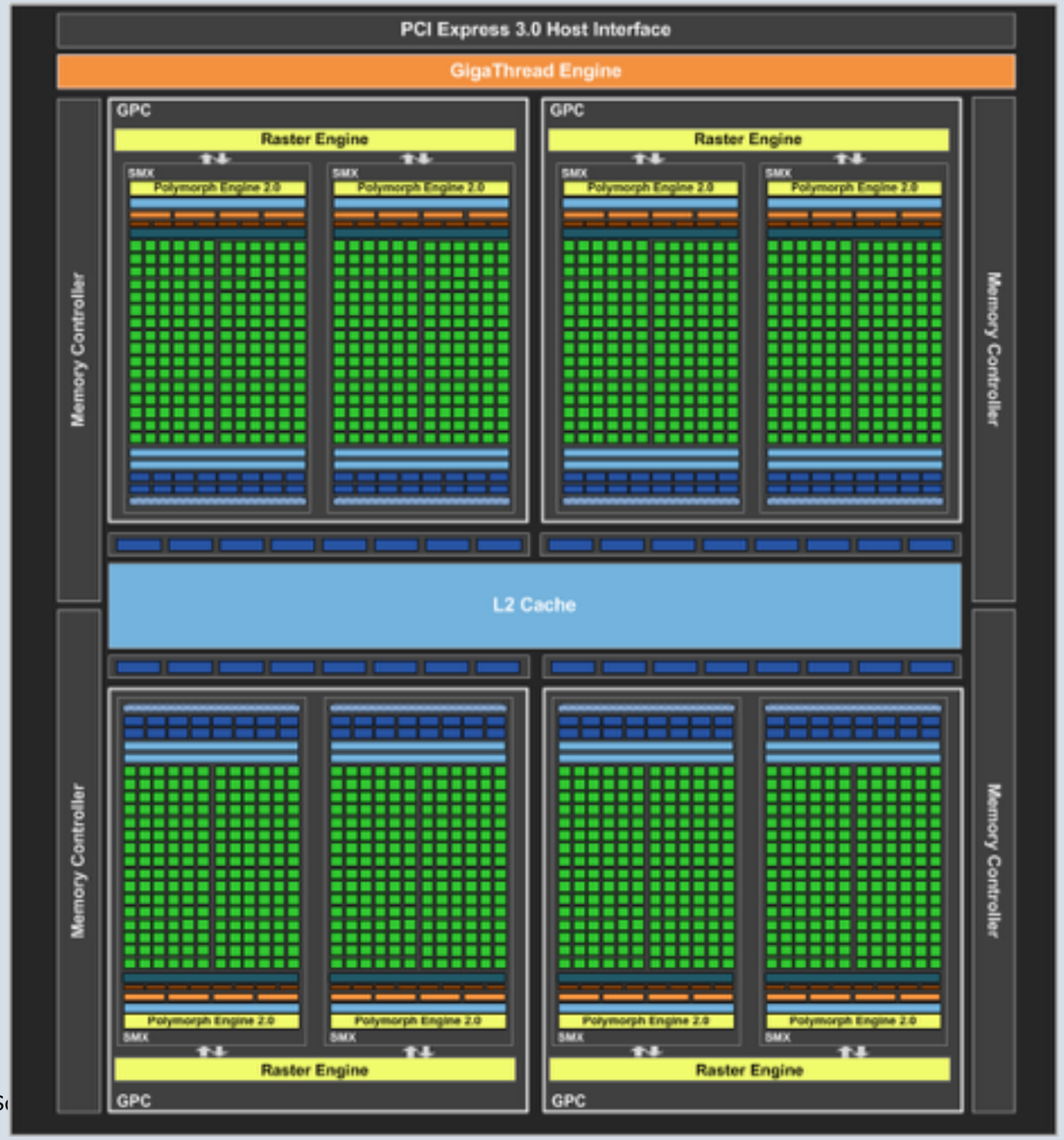
- **CUDA Sample: deviceQuery**

```
C:\ProgramData\NVIDIA Corporation\CUDA Samples\v6.5\bin\win64\Release>deviceQuery.exe
deviceQuery.exe Starting...

 CUDA Device Query (Runtime API) version (CUDART static linking)

Detected 1 CUDA Capable device(s)

Device 0: "GeForce GTX 680"
  CUDA Driver Version / Runtime Version          6.5 / 6.5
  CUDA Capability Major/Minor version number:    3.0
  Total amount of global memory:                 2048 MBytes (2147483648 bytes)
  ( 8) Multiprocessors, (192) CUDA Cores/MP:     1536 CUDA Cores
  GPU Clock rate:                                1059 MHz (1.06 GHz)
  Memory Clock rate:                             3004 Mhz
  Memory Bus Width:                              256-bit
  L2 Cache Size:                                 524288 bytes
  Maximum Texture Dimension Size (x,y,z)         1D=(65536), 2D=(65536, 65536), 3D=(4096, 4096, 4096)
  Maximum Layered 1D Texture Size, (num) layers  1D=(16384), 2048 layers
  Maximum Layered 2D Texture Size, (num) layers  2D=(16384, 16384), 2048 layers
  Total amount of constant memory:               65536 bytes
  Total amount of shared memory per block:       49152 bytes
  Total number of registers available per block: 65536
  Warp size:                                     32
  Maximum number of threads per multiprocessor:  2048
  Maximum number of threads per block:           1024
  Max dimension size of a thread block (x,y,z): (1024, 1024, 64)
  Max dimension size of a grid size    (x,y,z): (2147483647, 65535, 65535)
  Maximum memory pitch:                          2147483647 bytes
  Texture alignment:                             512 bytes
  Concurrent copy and kernel execution:          Yes with 1 copy engine(s)
  Run time limit on kernels:                     No
  Integrated GPU sharing Host Memory:            No
  Support host page-locked memory mapping:       Yes
  Alignment requirement for Surfaces:            Yes
  Device has ECC support:                        Disabled
```

Kepler Microarchitecture

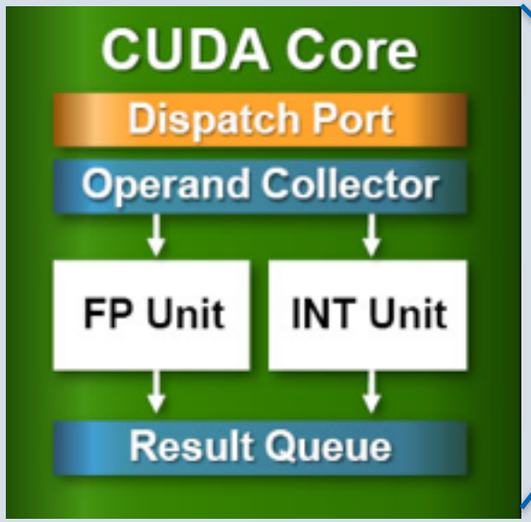
- **GTX 680**
  - ✓ Codenamed “GK104”
  - ✓ 3.54 billion transistors
  - ✓ 8 SMX
  - ✓ 1536 CUDA Cores
  - ✓ 3090 GFLOPs
  - ✓ 192GB/s Memory BW
  - ✓ TSMC’s 28nm manu.
  - ✓ TDP 195W

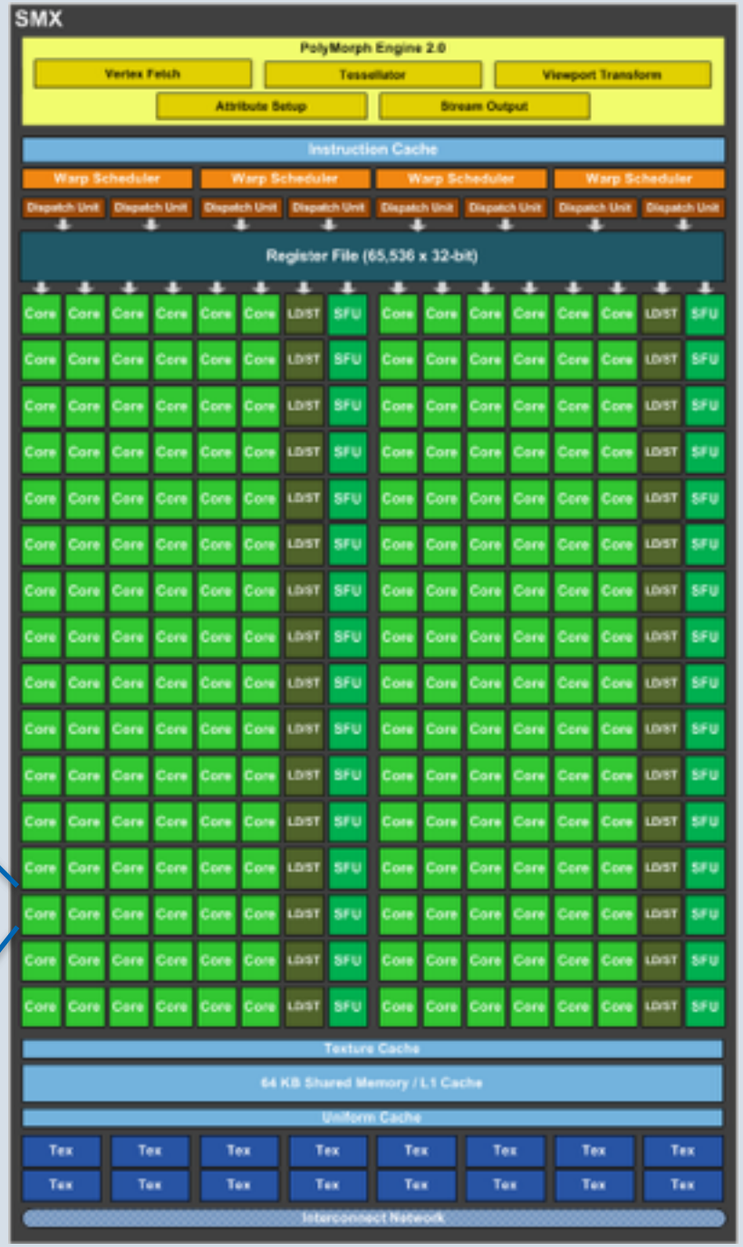
Refer to “NVIDIA GeForce GTX 680 Whitepaper”.



Kepler SMX Processor

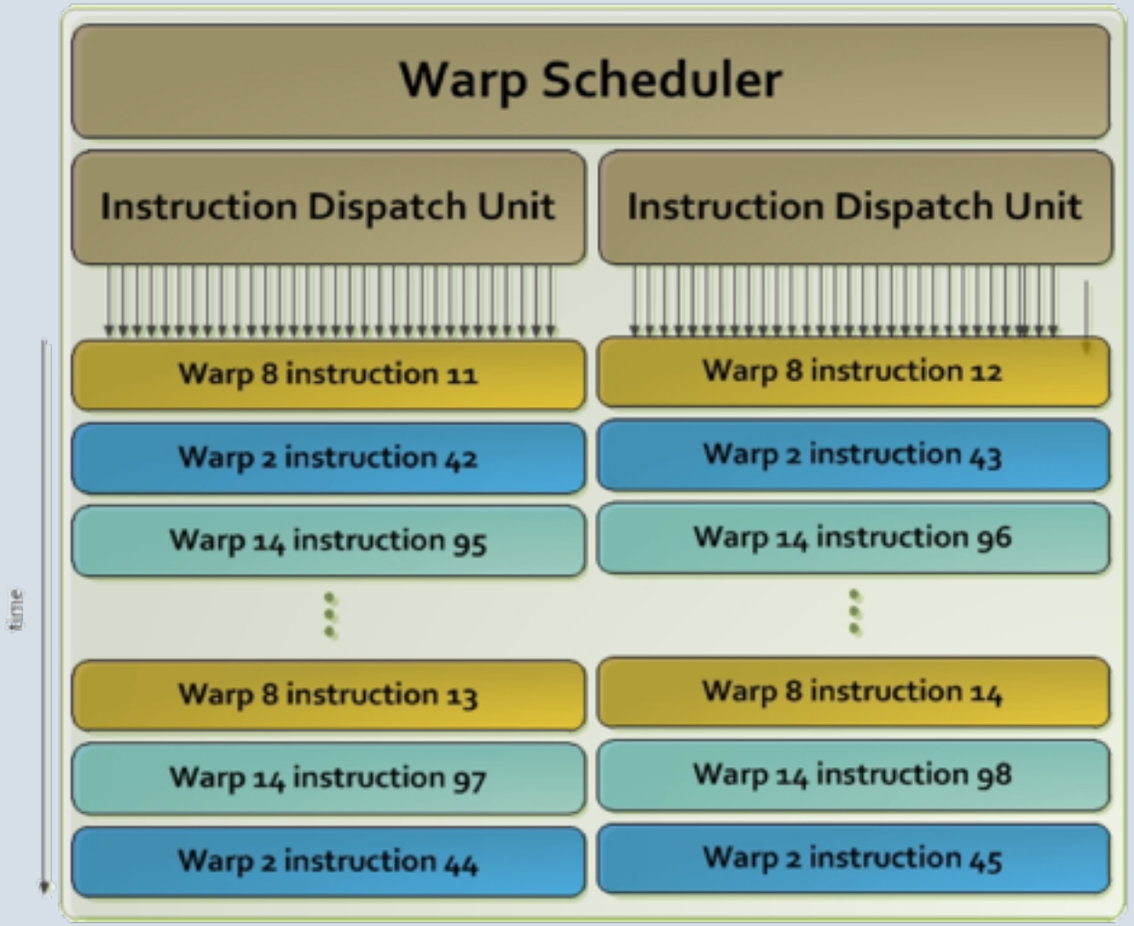
- **SMX (GK104)**
  - ✓ **192 CUDA Cores**
  - ✓ **Runs at graphics clock**
  - ✓ **32 LD/ST units**
  - ✓ **32 SFU**
  - ✓ **64KB shared mem/L1 Cache**
  - ✓ **64K Registers**
  - ✓ **4 Warp Schedulers**





Kepler Quad-Warp Scheduler

- **Warp Scheduler**
  - ✓ **1 Warp = 32 parallel threads**
  - ✓ **each warp scheduler is capable of dispatching two instructions per warp every clock**
  - ✓ **Each warp allows two independent instructions per cycle**



Kepler

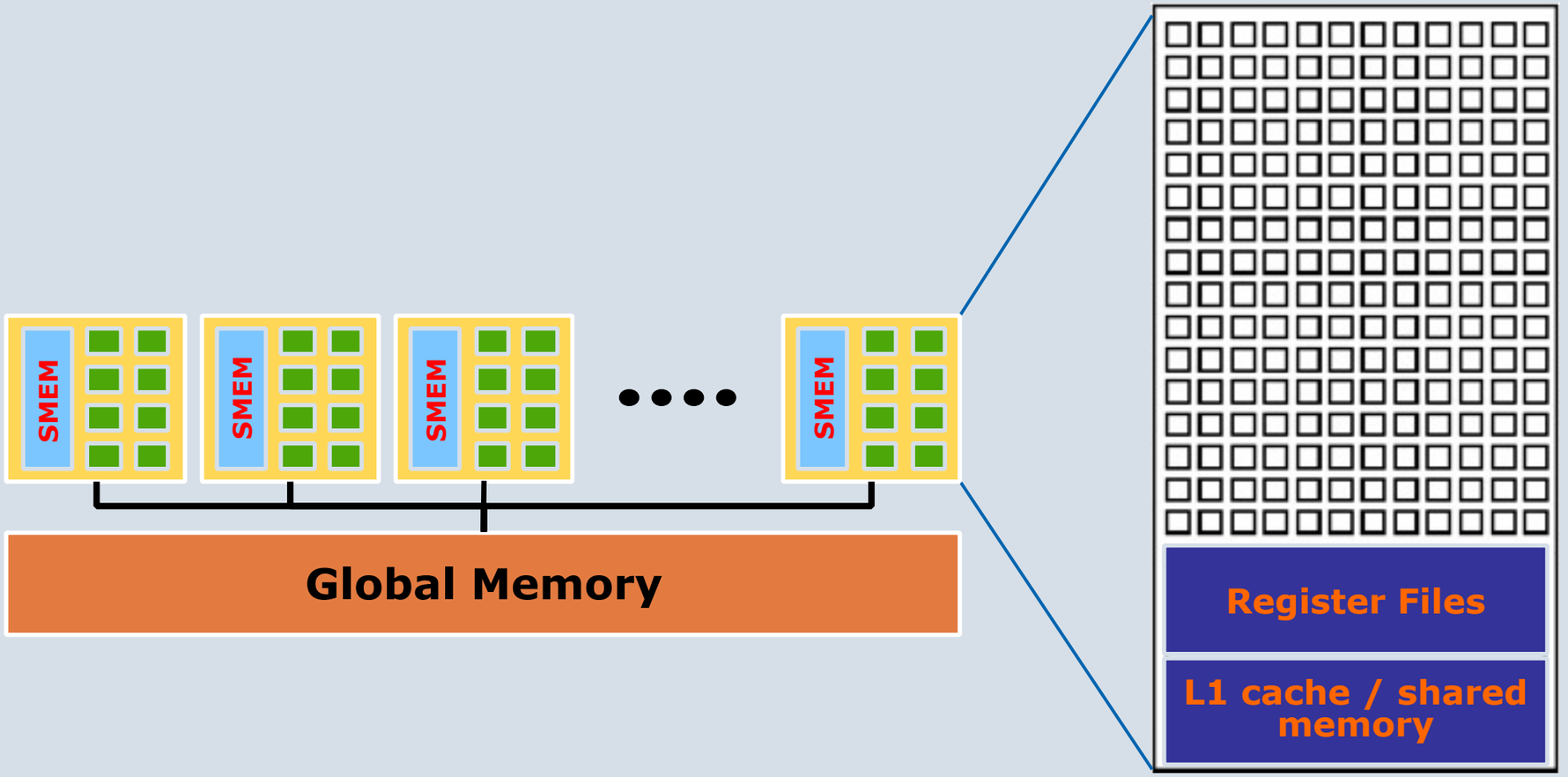
- There are multiple products in the latest Kepler generation
- Consumer graphics cards (GeForce):
  - ✓ GTX650 Ti: 768 cores, 1/2GB
  - ✓ GTX660 Ti: 1344 cores, 2GB
  - ✓ GTX680: 1536 cores, 2/4GB
  - ✓ GTX690: 2×1536 cores, 2×2GB
  - ✓ GTX 780: 2304 cores, 3GB
  - ✓ GTX TITAN: 2688 cores, 6GB
- HPC cards (Tesla):
  - ✓ K10 module: 2×1536 cores, 2×4GB
  - ✓ K20 card: 2496 cores, 5GB
  - ✓ K20X module: 2688 cores, 6GB
  - ✓ K40 card:2880 cores, 12GB
  - ✓ K80 module:4992 cores, 24GB

Kepler

- Building block is a “streaming multiprocessor” (SMX):
  - ✓ 192 cores and 64k registers
  - ✓ 64KB of shared memory / L1 cache
  - ✓ 8KB cache for constants
  - ✓ 48KB texture cache for read-only arrays
  - ✓ up to 2K threads per SMX
- Different chips have different numbers of these SMXs:

product	SMXs	bandwidth	memory	power
GTX 650 Ti	4	86 GB/s	1/2 GB	110W
GTX 680	8	190 GB/s	2/4 GB	195W
K10 (2×)	8	160 GB/s	4 GB	110W
K20X	14	250 GB/s	6 GB	235W

Kepler SMX


SMEM
SMEM
SMEM
SMEM
Global Memory
Register Files
L1 cache / shared memory

Quiz

- **Maxwell架构是Kepler架构的下一代产品，请通过查阅白皮书等资料**
  - ✓ **总结出Maxwell架构对Kepler架构的改进之处**

Outline

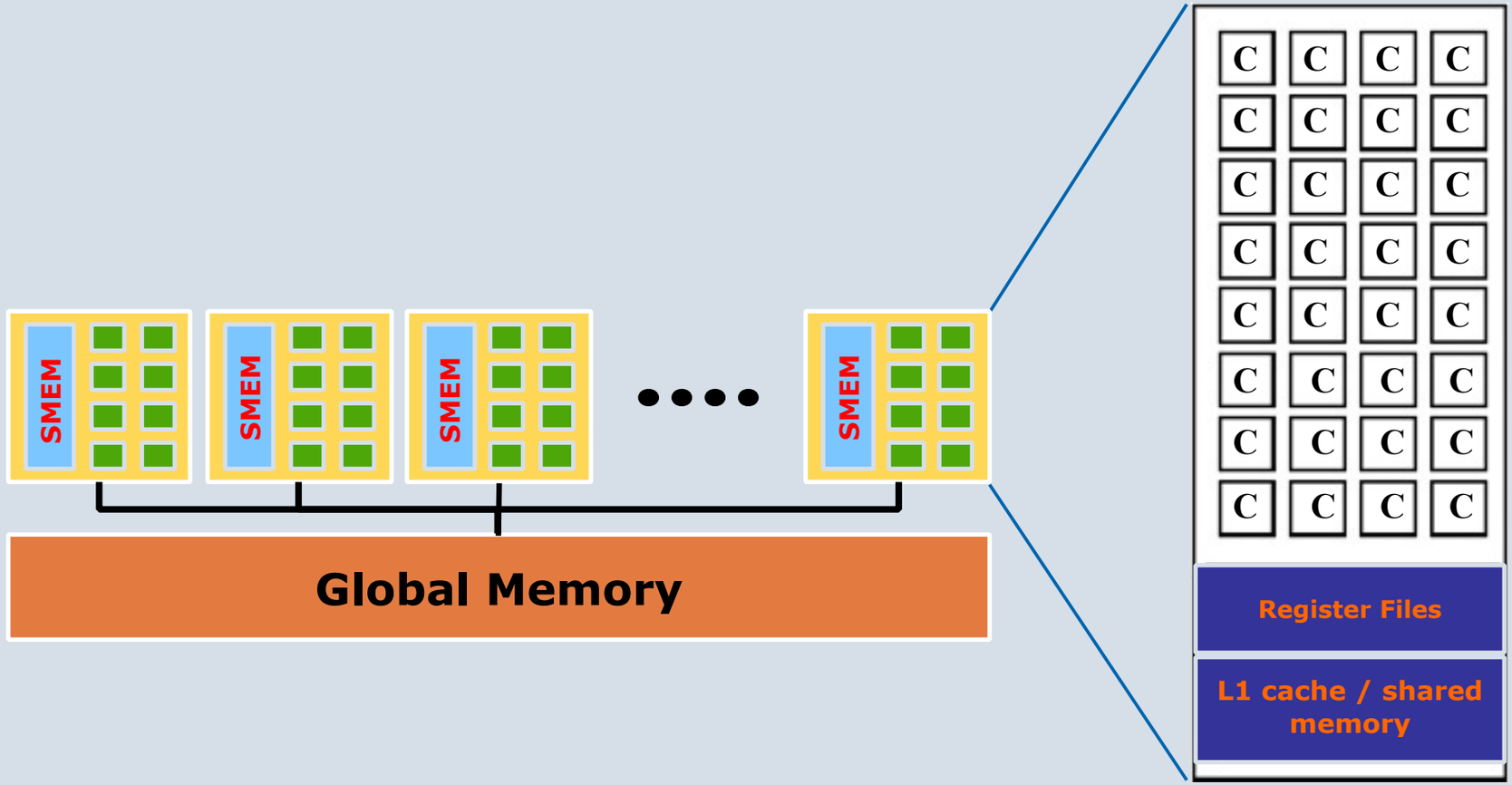
1. Linking Model
2. Kepler Architecture
3. Fermi Architecture

Fermi

- Older Fermi GPU has SM （“Streaming multiprocessor”）:
  - ✓ 32 cores and 32k registers
  - ✓ 64KB of shared memory / L1 cache
  - ✓ 8KB cache for constants
  - ✓ up to 1536 threads per SM
- Different chips have different numbers of these SMs:

product	SMs	bandwidth	memory
GTX 560	14	130 GB/s	1/2 GB
GTX 580	16	190 GB/s	1.5 GB
M2050/2070	14	140 GB/s	3/6 GB
M2075/2090	16	140 GB/s	3/6 GB

Fermi SM


SMEM
SMEM
SMEM
SMEM
Global Memory
C
Register Files
L1 cache / shared memory

Quiz

- **Fermi架构是Kepler架构的上一代产品**
  - ✓ 请对比**Fermi**架构与**Kepler**架构在设计上的区别