Many-core Architectures

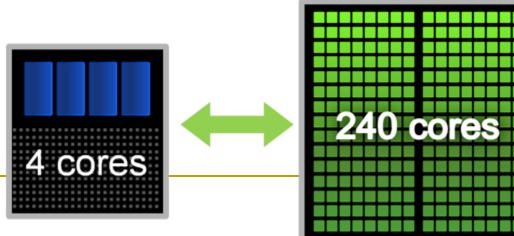
Sathish Vadhiyar Introduction to Scalable Systems

Introduction

- Graphical Processing Unit
- A single CPU can consist of a max of 128 cores
- GPUs consist of a large number of lightweight cores
- Primarily proposed for graphics applications
 - Consists of stages with each stage consisting of independent computations

GPU and CPU, CUDA

- Typically GPU and CPU coexist in a heterogeneous setting
- "Less" computationally intensive part runs on CPU (coarse-grained parallelism), and more intensive parts run on GPU (fine-grained parallelism)
- NVIDIA's GPU architecture is called CUDA (Compute Unified Device Architecture) architecture, accompanied by CUDA programming model, and CUDA C language

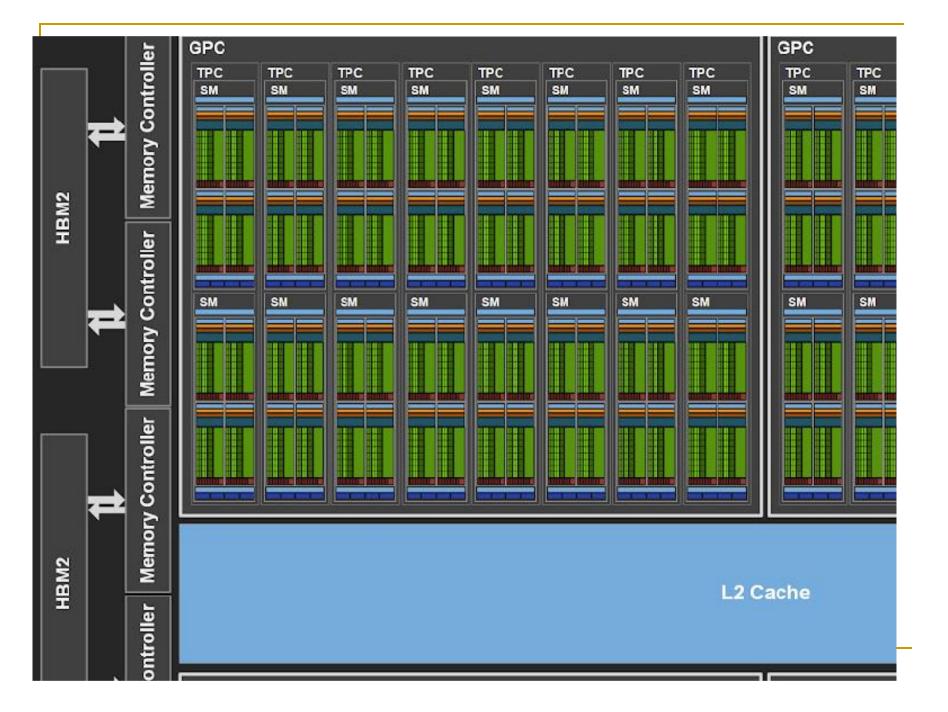


NVIDIA A100

- Graphic Processing Clusters (GPCs) 7 Nos.
- Each GPC has:
 - 7 or 8 Nos. Texture Processing Clusters (TPCs).
 Hence a total of 64 TPCs
 - Up to 16 Nos. Streaming Multiprocessors (SMs).
 A total of 108 SMs



Figure 6. GA100 Full GPU with 128 SMs (A100 Tensor Core GPU has 108 SMs)



NVIDIA A100 SM

Each SM has:

- 64 Nos. FP32 GPU cores. A total of 6912 GPU cores.
- 4 tensor cores. Hence a total of 432 tensor cores. Together they deliver 1024 FP16/FP32 operations per clock
- 192 KB of combined shared memory and L1 data cache

SM

		_								ion Cac											
					tion C					L0 Instruction Cache											
						hread/c				Warp Scheduler (32 thread/clk)											
		Di	spatch	Unit	(32 th	read/cli	k)			Dispatch Unit (32 thread/clk)											
		Reg	jister i	File (1	16,384	4 x 32-	Ыt)					Reg	ister	File (16,38	4 x 32	2-bit)				
INT32 I	NT 32	FP32	FP32	FP	64					INT32	INT32	FP32	FP32	FF	164						
INT32 II	NT32	FP32	FP32	FP	64					INTER	INT32	FP32	FP32		164						
INT32 I	NT32	FP32	FP32	FP	64					INTER	INT32	FP32	FP32	FF	64						
INTER	NT32	FP32	FP32	FP	64		TENSOR CORE			INTER	INT32	FP32	FP32	FF	64	TENSOR CORE					
ютаа и	NT32	FP32	FP32	FP	64	I EI				інтаа	INT32	FP32	FP32	FF	¹⁶⁴				ince		
INT32 I	NT32	FP32	FP32	FP	64					INT32	INT32	FP32	FP32	FI	64						
INT32	NT32	FP32	FP32	FP	64					INTER	INT32	FP32	FP32	FF	164						
INT32	NT32	FP32	FP32	FP	64					інтаа	INT32	FP32	FP32	FF	×64						
UD/ 81		110/ 81	UD/ 81	LDV ST	110/ 81	LDV 8T	LD/ 81	SFU		LD/ 81		UD/ 81	LDV 8T		LD/ 81	LDV ST	LDI 8T	s	FU		
	_									_											
	L0 Instruction Cache											L0 Instruction Cache									
_	_					hread/cli				Warp Scheduler (32 thread/clk) Dispatch Unit (32 thread/clk)											
			sparen	unit	(3 2 m	real serveral	~1		- 11	Dispatch Unit (32 threadvolk)											
		Reg	jister f	File (1	16,384	1 x 32-	bit)					Reg	ister	File (16,38	4 x 31	2-bit)				
INT32 I	NT32	FP32	FP32	FP	64					INT32	INT32	FP32	FP32	FT	64						
INT32 #	NT32	FP32	FP32	FP	64					INTER	INT32	FP32	FP32		64						
INTER	NT32	FP32	FP32	FP	64					інтаа	INT32	FP32	FP32	FF	×64						
INTEE	NT32	FP32	FP32	FP	64	TE	NRO	R CORE		інтаа	INT32	FP32	FP32	FF	P64		INSO		DE		
INT32 I	NT32	FP32	FP32	FP	64	1.2.1	1.50	C COME		INT32	INT32	FP32	FP32	FF	64		=14-300				
INT32 I	NT32	FP32	FP32	FP	64					INTS2	INT32	FP32	FP32		64						
INT32 I	NT32	FP32	FP32	FP	64					INTER	INT32	FP32	FP32	FF	164						
INT321	NT32	FP32	FP32	FP	64					інтаз	INT32	FP32	FP32	FF	964						
LDV	LDY 5T	LD/ 51	LD/ 51	LDY ST	LD/ 51	LDV ST	LD/ 51	SFU		LD/ 81	LDY 8T	LD/ 81	LDV 8T	LDF ST	LD/ 81	UDV 8T	1101 81	s	FU		
51								3 L1 Data	Carl												
						1	92KI	B C1 Data	Gaun	e/ana	red M	emory									



	L0 Instruction Cache											L0 Instruction Cache											
	Warp Scheduler (32 thread/clk)											Warp Scheduler (32 thread/clk)											
	Dispatch Unit (32 thread/clk)											Dispatch Unit (32 thread/clk)											
	Register File (16,384 x 32-bit)										Register File (16,384 x 32-bit)												
INT32	NT32	FP32	FP32	FP	64					INT32	INT32	FP32	FP32	FF	P64								
INT32	NT32	FP32	FP32	FP	64					INT32	INT32	FP32	FP32	FF	°64	TENSOR CORE							
INT32 I	NT32	FP32	FP32	FP	64					INT32	INT32	FP32	FP32	FF	P64								
INT32 I	NT32	FP32	FP32	FP	64	те		R CORE		INT32	INT32	FP32	FP32	FF	P64								
INT32	NT32	FP32	FP32	FP	64	TENSOR CORE				INT32	INT32	FP32	FP32	32 FP64				N CONL					
INT32	NT32	FP32	FP32	FP	64					INT32	INT32	FP32	P32 FP32 FP		P64								
INT32	NT32	FP32	FP32	FP	64					INT32	INT32	FP32	FP32	FF	P64								
INT32	NT32	FP32	FP32	FP	64					INT32	INT32	FP32	FP32	FF	FP64								
LD/ ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	SFU		LD/ ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	LD/ ST	SFU					
							192K	B L1 Data C	ach	e / Sha	red M	emory	1										
	Tex Tex											Тех					Tex	ĸ					

Tensor cores

- Specialized HPC compute cores for matrix operations for AI and HPC applications
- Perform matrix multiply and accumulate (MMA) operations

Streaming Multiprocessor (SMX) Architecture in K40

- 65536 registers
- 192 SP cores
- 64 DP cores
- 32 Special functional units (SFUs)
- 32 load/store units
- 16 texture filtering units

								Ins	tructi	on Ca	che										
	War	p Sch	neduler			Wa	rp Scher	Scheduler Warp Se					veduler			Warp Scheduler					
Dispatch Dispatch					Dispatch Dispatch						spate	h	Dispat	Dispatch Dispatch							
							Regi	ster i	File ((65,53	6 x 3	2-bit)									
÷	÷	٠	+	+	+	+	+	+	+	+	٠		+		+		+	+	4		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LOIST	SF		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LDIST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LOIST	sr		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LDIST	sı		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LDIST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LOST	SF		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LOIST	sr		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/BT	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LOIST	sr		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LDIST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LOST	SF		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LDIST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LOIST	si		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/8T	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LOIST	sr		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LDIST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LOST	si		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/8T	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LDIST	si		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LDIST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LOST	si		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LDST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LOIST	si		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LD/ST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LDIST	SF		
Core	Core	Core	DP Unit	Core		Core	DP Unit	LDST	SFU	Core			DP Unit				DP Unit	LOST	SF		
Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LINST	SFU	Core	Core	Core	DP Unit	Core	Core	Core	DP Unit	LOIST	SI		
										ct Net											
							64 KB					Cac	he								
							48 K	B Re	ad-O	nly D	ata C	ache)								
	Tex		Tex	:	Tex			Tex	¢	Тех			Tex	Tex			Tex				
	Tex		Tex			Tex		Tex			Tex		Tex			Tex		Tex			

Shared memory size = 16/32/48 K

CUDA Memory Spaces

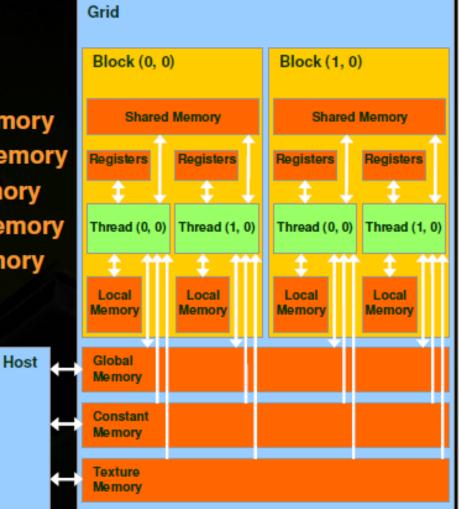


JI

Each thread can:

- Read/write per-thread registers
- Read/write per-thread local memory
- Read/write per-block shared memory
- Read/write per-grid global memory
- Read only per-grid constant memory
- Read only per-grid texture memory

The host can read/write global, constant, and texture memory (stored in DRAM)



Memory Hierarchy

- Global or Device memory:
 - Can be accessed by all the threads executing in all the SMXs
 - Can be accessed by the CPU host
 - □ Kepler K40: 12 GB.
 - In A100s: 80 GB
- Shared memory:
 - In each SMX
 - Shared by all the threads of a thread block executing in a SMX
 - Kepler K40: 64 KB can be configured as 16/32/48KB for shared memory, with the rest for L1 cache
 - In A100s: 192 KB

Memory Hierarchy

- Registers
 - In each SMX
 - Used for storing the local data of the threads
 - Kepler k40: 64K registers in each SMX
 - A100: 64K registers in each SMX
- Constant and texture memory
 - Accessed by all the threads
 - Texture memory: used to improve performance of reads that exhibit spatial locality among the threads
- Latency of data access
 - Device memory: 200-400 clock cycles (about 300 ns)
 - Shared memory: 20-30 clock cycles (about 5 ns)

Differences with the CPU threads

- Context switching fast: The state of a thread (thread block) stored in shared memory and registers stay till execution completion
- Cache explicitly managed: User's program will have to explicitly bring the frequently accessed data from the device to the shared memory

Backup

SPs and SMX

- GPU cores are called streaming processors (SPs)
- SP cores grouped into streaming multiprocessors (SMX)
- Promotes coarse and fine-level parallelism
- Kepler K40 15 SMXs consisting of 192 SP cores each for a total of 2880 SP cores
- Light-weight 1.065 GHz
- SIMD execution (actually SIMT)

