NVIDIA CUDA И OPENACC
ЛЕКЦИЯ 1

Перепёлкин Евгений

СОДЕРЖАНИЕ

Лекция 1

- Введение
- Гибридная модель вычислений
- Типы вычислительных архитектур
- Архитектура графического процессора GPU

Введение

ЭВОЛЮЦИЯ ЦЕНТРАЛЬНЫХ ПРОЦЕССОРОВ

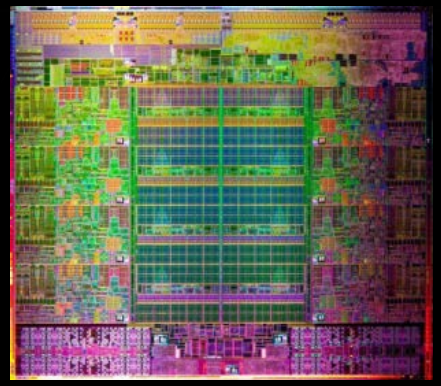
Частоты CPU

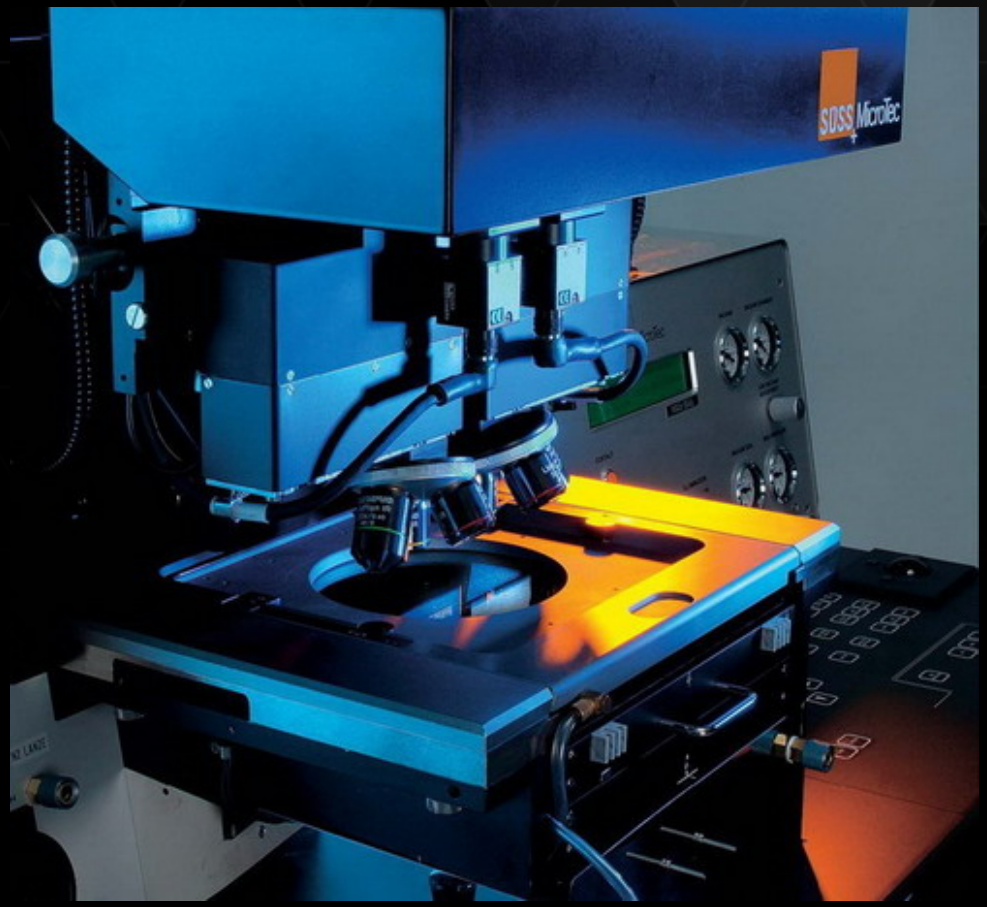
- 2004 г. Pentium 4, 3.46 GHz
- 2005 г. Pentium 4, 3.8 GHz
- 2006 г. Core Duo T2700, 2.33 GHz
- 2007 г. Core 2 Duo E6700, 2.66 GHz
- 2007 г. Core 2 Duo E6800, 3 GHz
- 2008 г. Core 2 Duo E8600, 3.33 GHz
- 2009 г. Core i7 950, 3.06 GHz (4 ядра)
- 2010 г. Core i7 9xxX, 3.2 - 3.47 GHz (6 ядер)

ПРОБЛЕМЫ

Рост частоты практически отсутствует

- Энерговыделение ~ второй степени частоты
- Ограничения техпроцесса

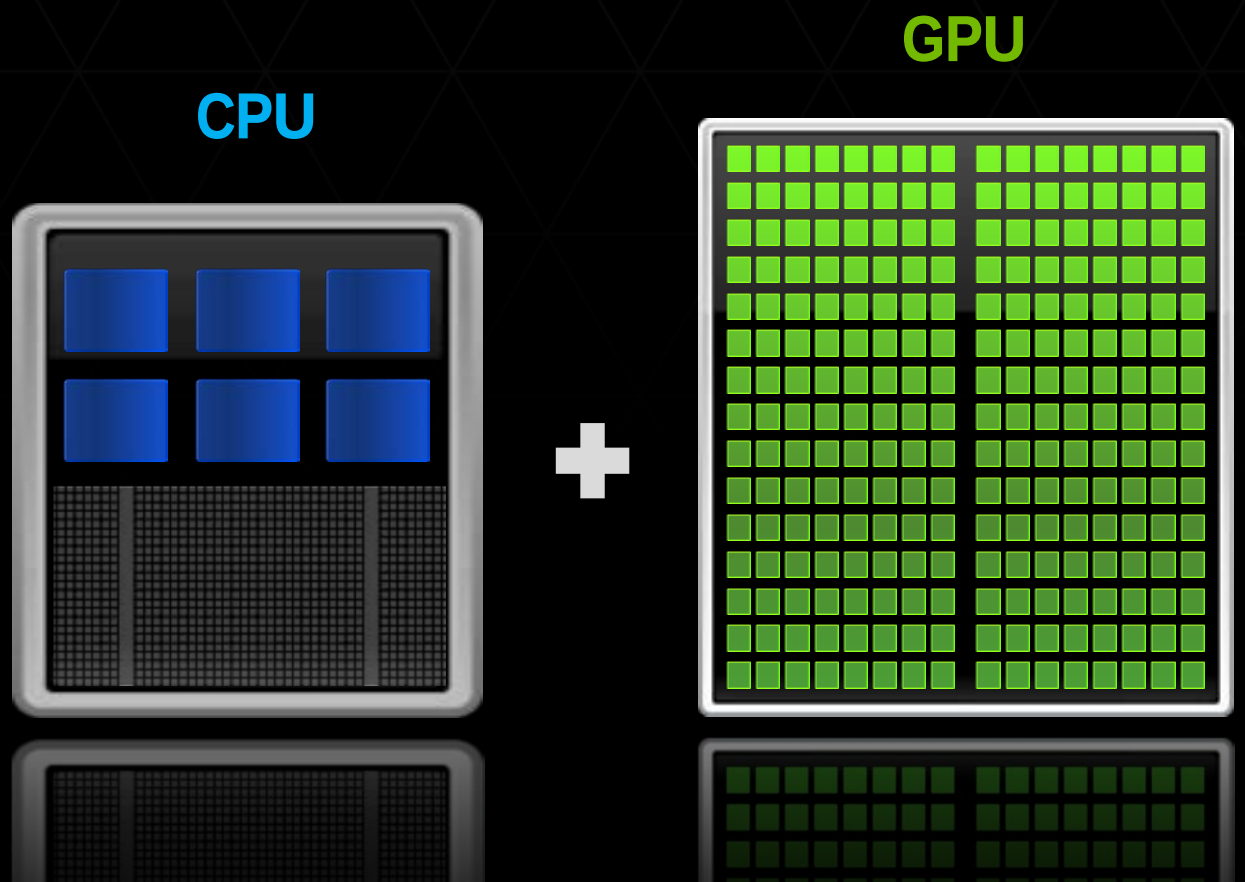




Гибридная модель вычислений

РАБОТА В ТАНДЕМЕ

Распределение вычислительной нагрузки между архитектурами



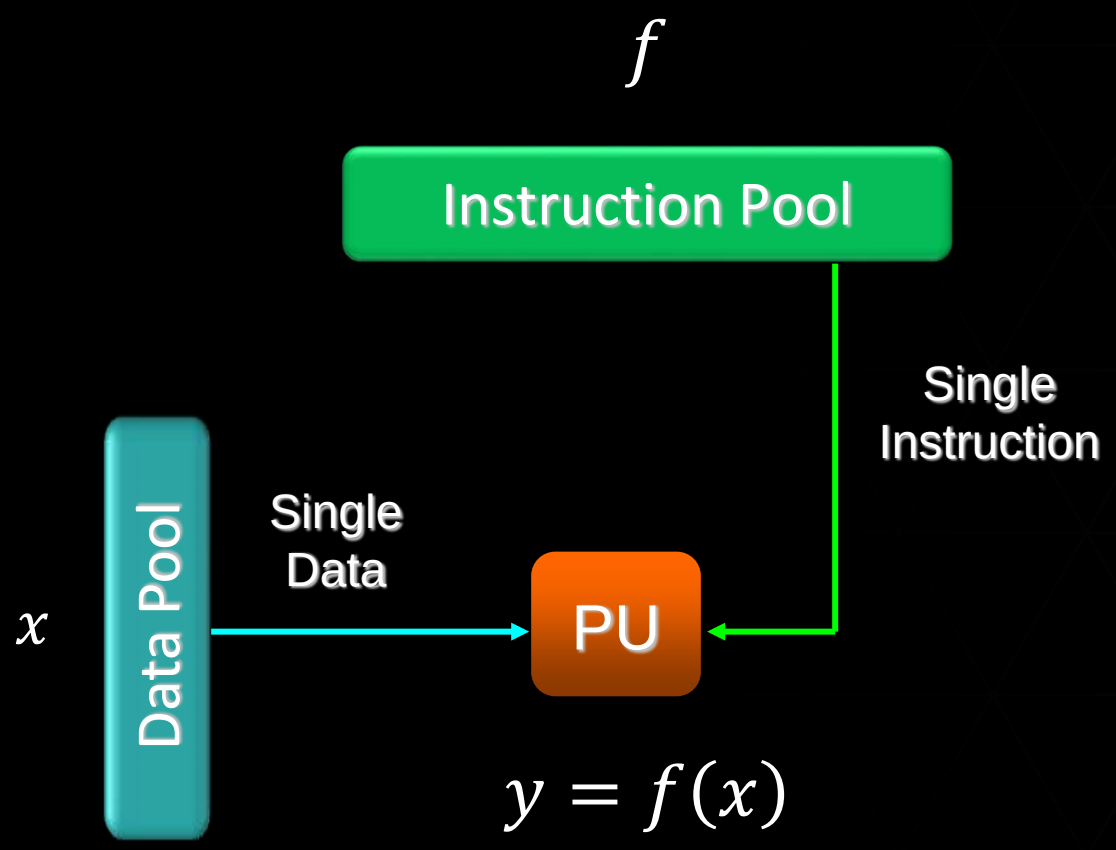
ГИБРИДНАЯ МОДЕЛЬ ВЫЧИСЛЕНИЙ

Типы вычислительных архитектур

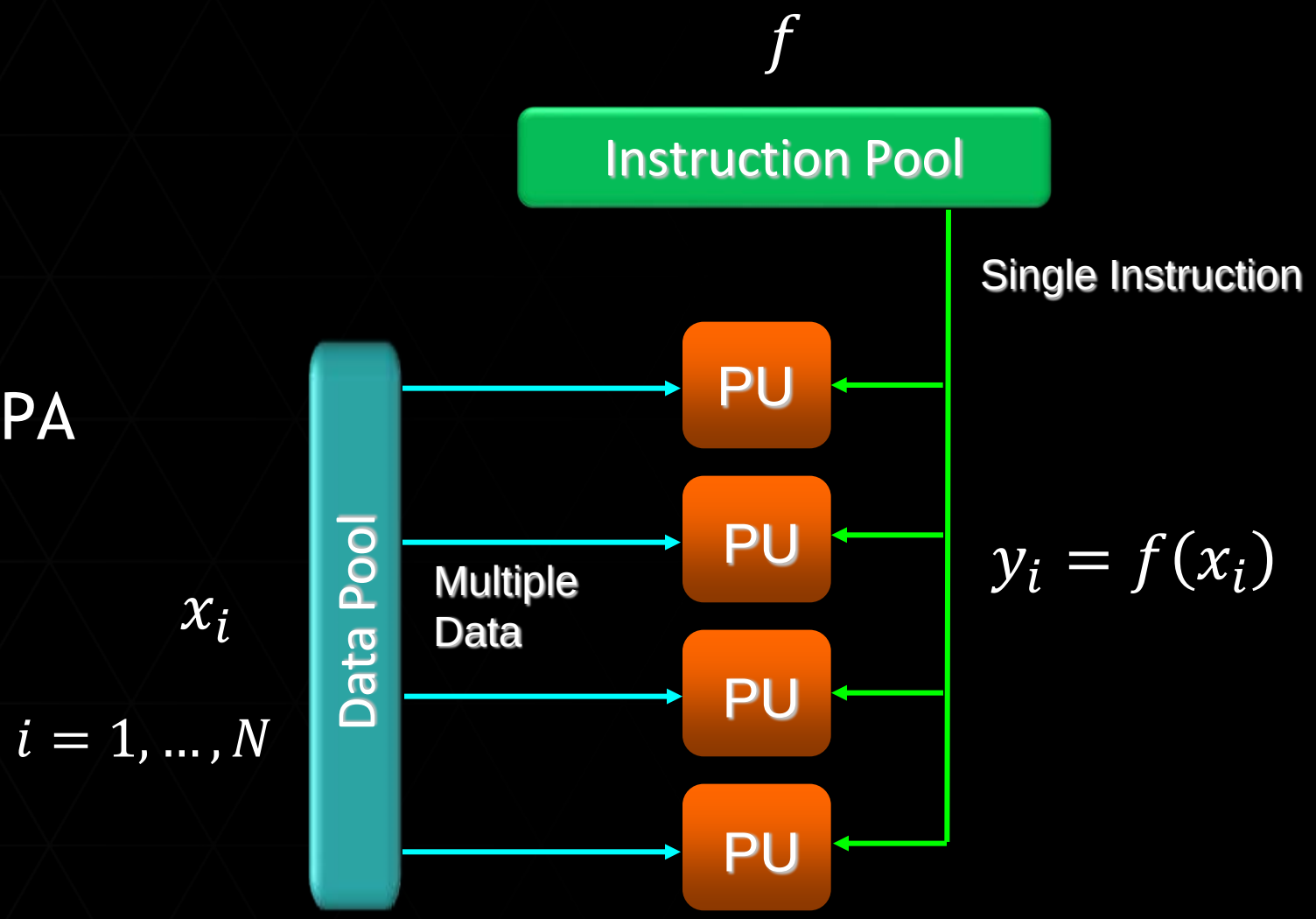
КЛАССИФИКАЦИЯ

	Single Instruction	Multiple Instruction
Single Data	SISD	MISD
Multiple Data	SIMD	MIMD

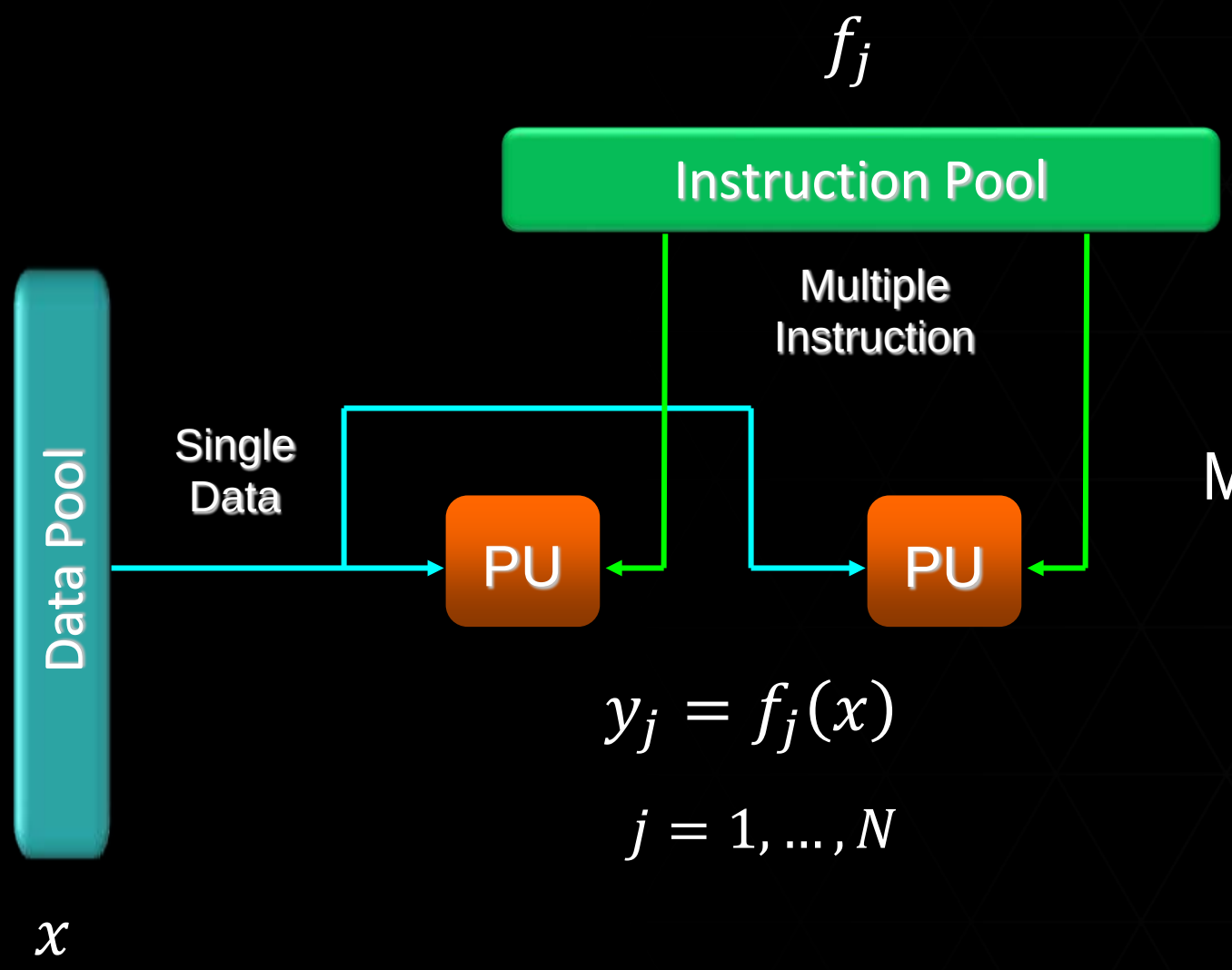
SISD АРХИТЕКТУРА



SIMD АРХИТЕКТУРА



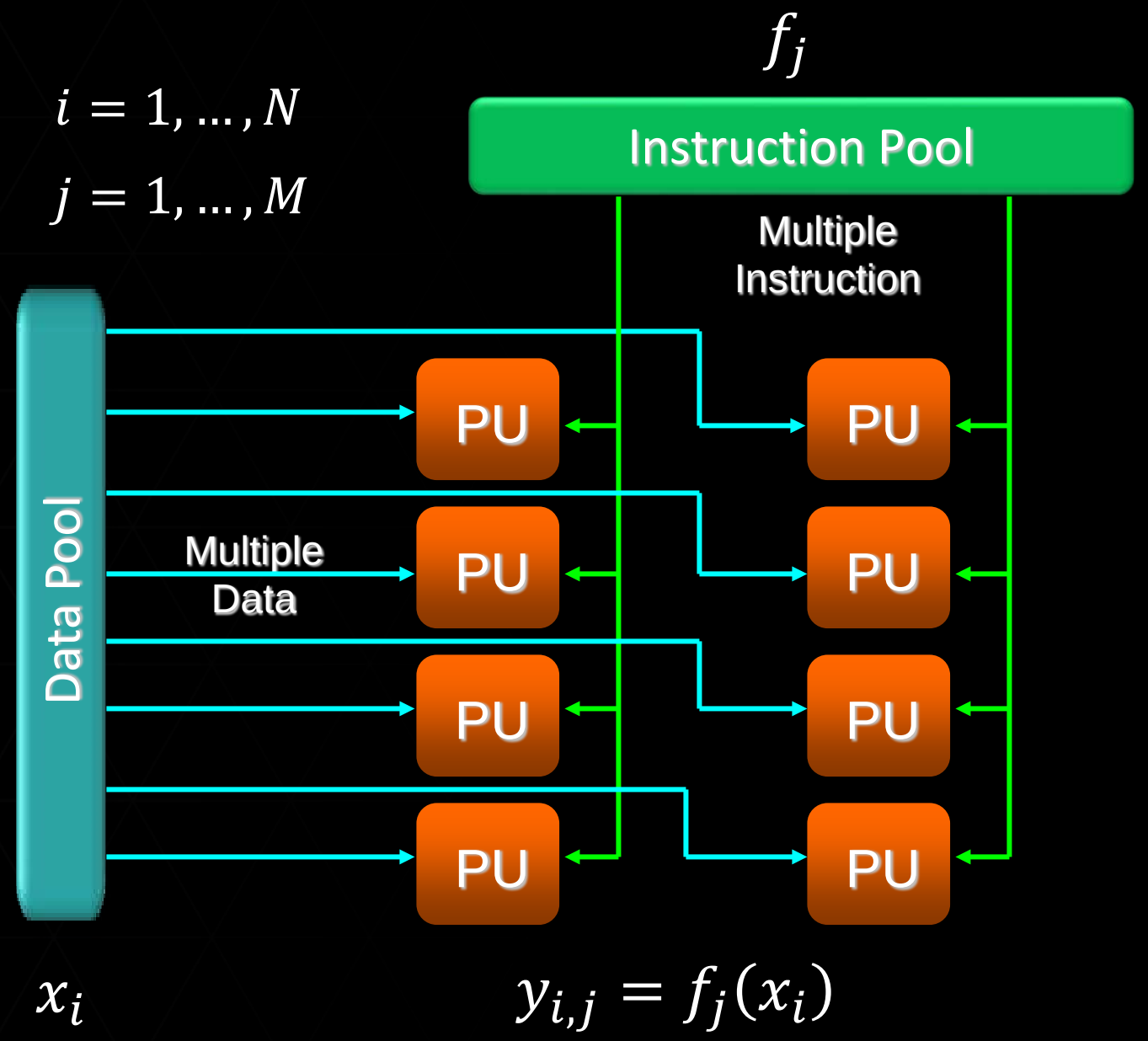
MISD АРХИТЕКТУРА



$$y_j = f_j(x)$$

$$j = 1, \dots, N$$

MIMD АРХИТЕКТУРА

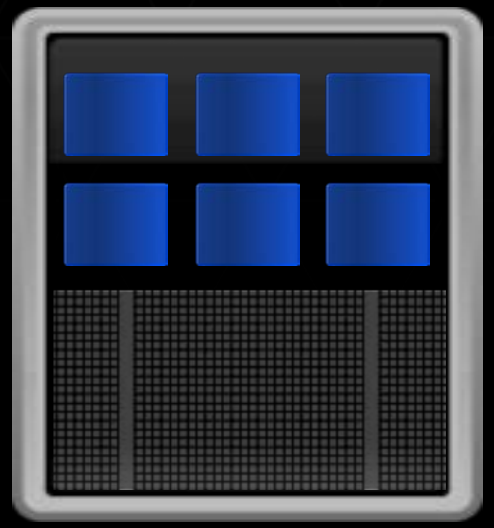


Архитектура графического процессора GPU

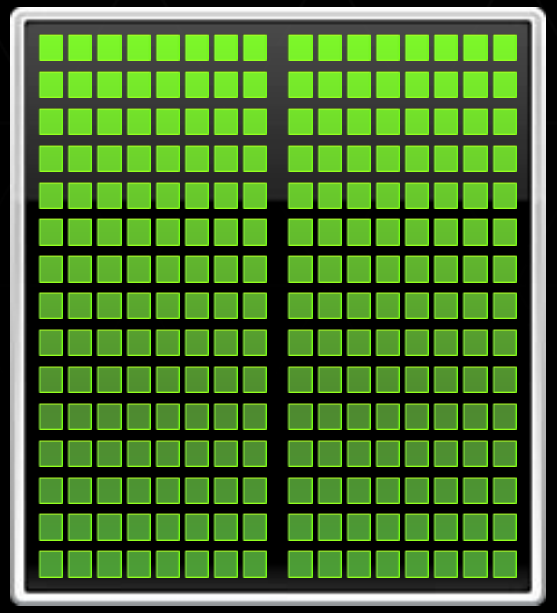
ОСОБЕННОСТИ АРХИТЕКТУР

Сравнение

CPU



GPU



WORLD LEADER IN VISUAL COMPUTING

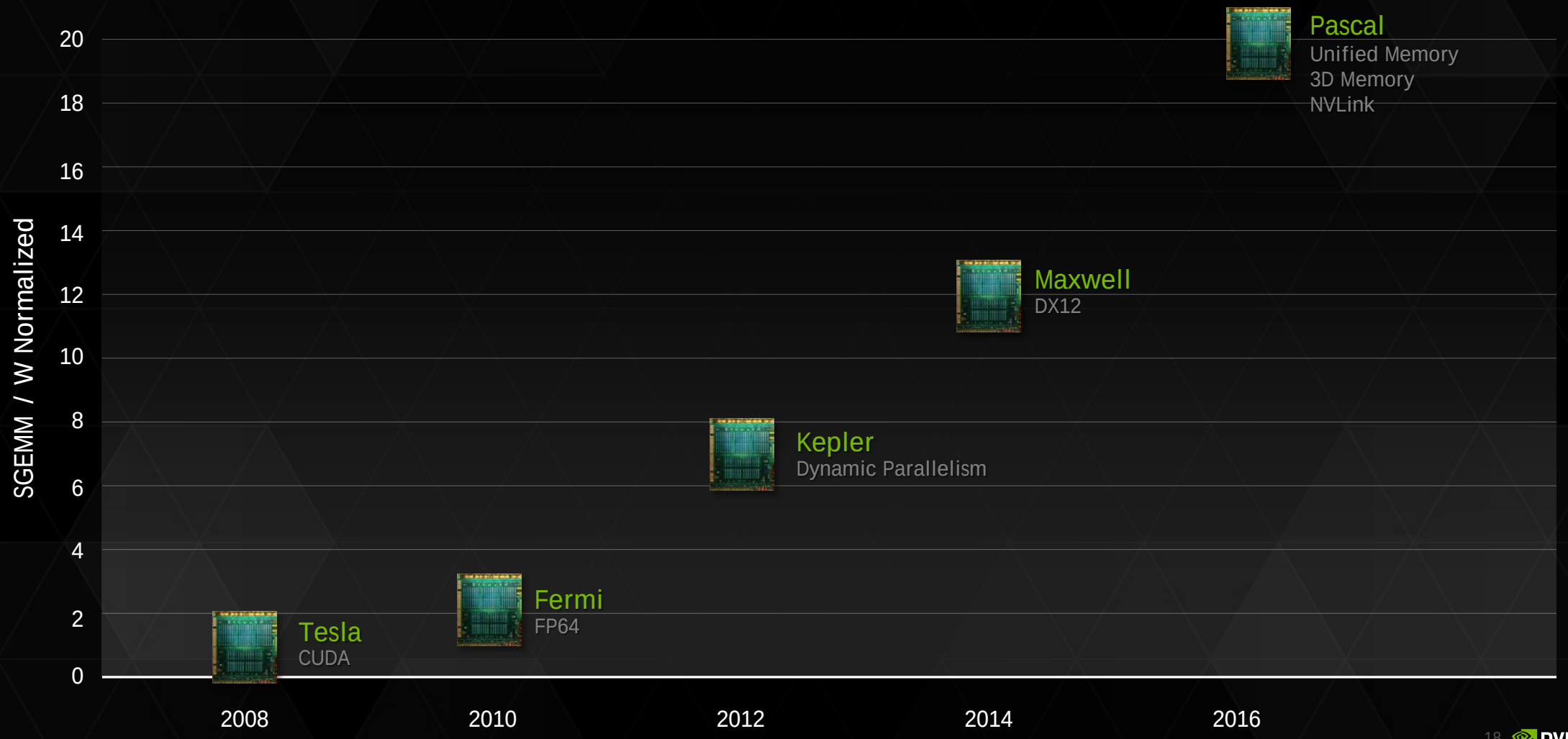
GAMING

PRO VISUALIZATION

HPC & BIG DATA

MOBILE COMPUTING

РАЗВИТИЕ GPU АРХИТЕКТУРЫ



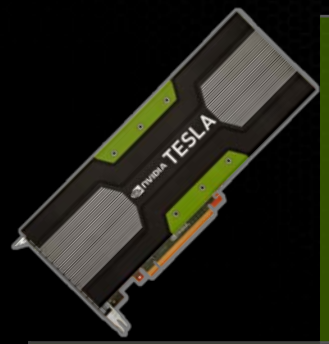
NVIDIA TESLA

Tesla C

Tesla M

Tesla S

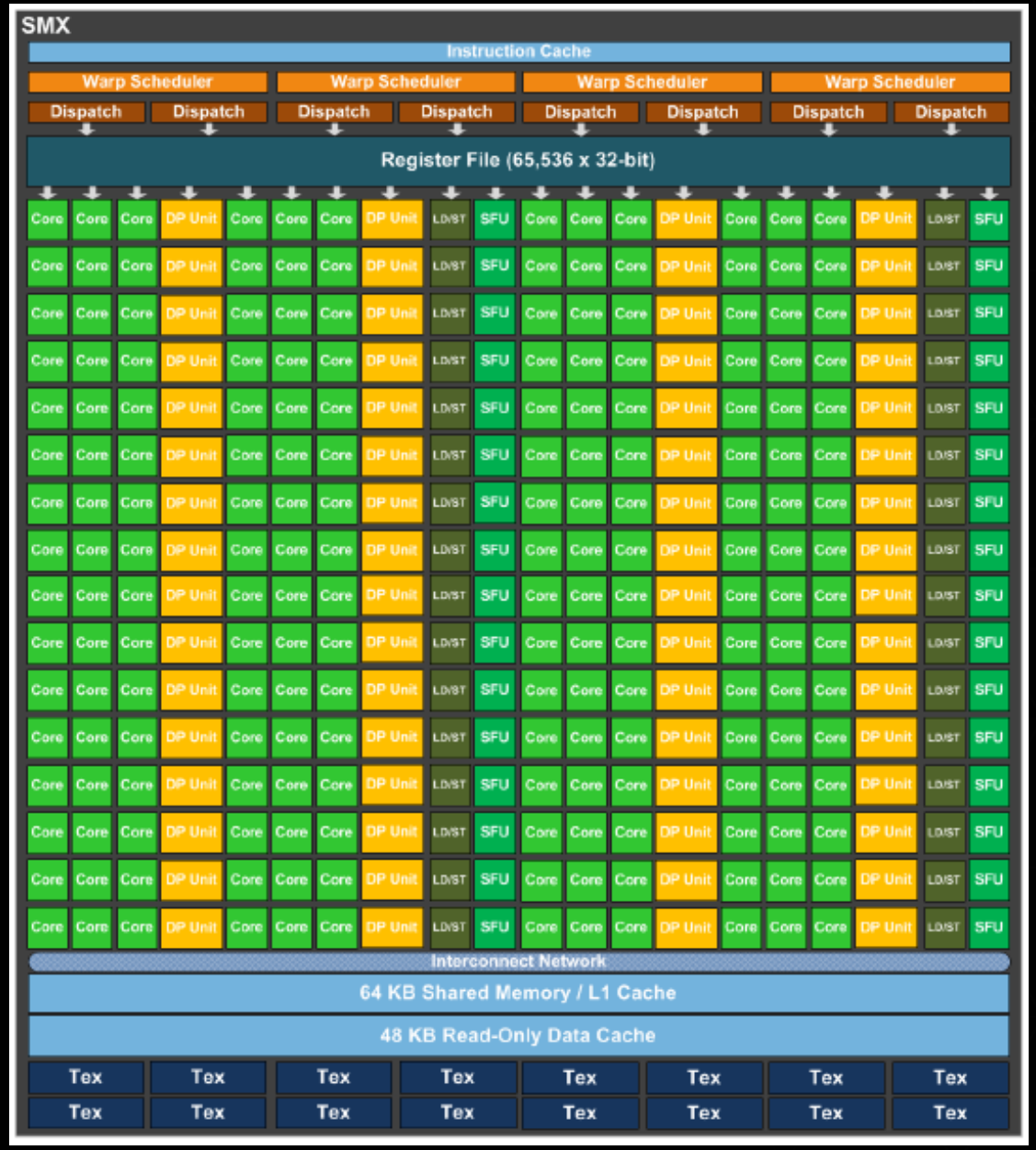
СЕМЕЙСТВО TESLA



	SM/TPC	FP32 CUDA cores SM/GPU	FP64 CUDA cores SM/GPU	Base clock, MHz	Peak, GFLOP FP32/64	Memory interface	Memory size, GB	Register file SM/GPU KB	Interface
Tesla K40 GK 110 (Kepler)	15 15	192 2880	64 960	745	5040 1680	384-bit GDDR5	Up to 12	256 3840	PCIe (Gen3)
Tesla M40 GM 200 (Maxwell)	24 24	128 3072	4 96	948	6840 210	384-bit GDDR5	Up to 24	256 6144	PCIe (Gen3)
Tesla P100 GP 100 (Pascal)	56 28	64 3584	32 1792	1328	10600 5300	4096-bit HBM2	16	256 14336	PCIe (Gen3) NVLink (up to 80 GB/s)

АРХИТЕКТУРА KEPLER GK 110

- 12 GB global memory
- 1.5 MB L2 Cache



СТРУКТУРА SMX

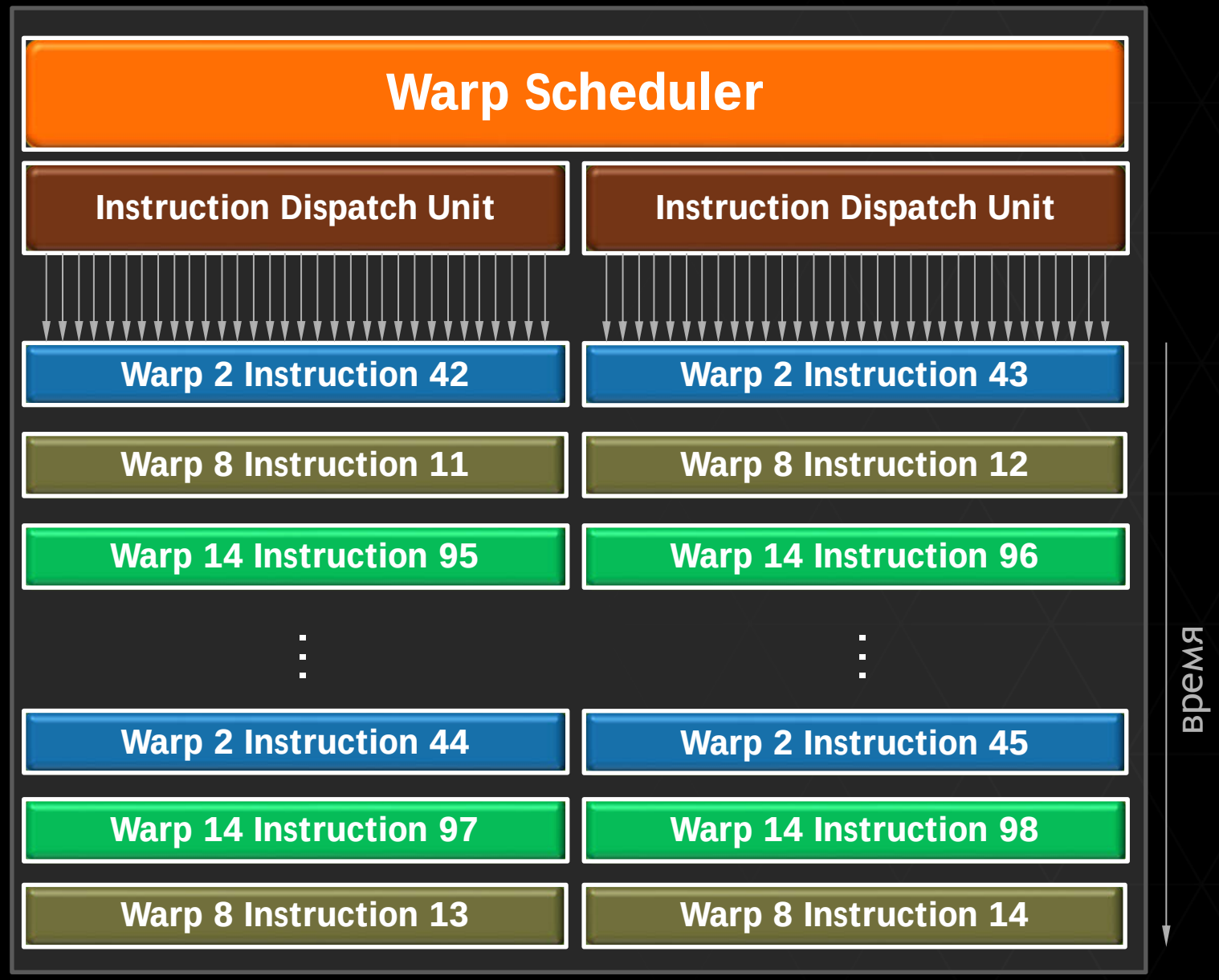
Core - 192 ядра

DP - 64 блока для вычислений с двойной точностью

SFU - 32 блока для вычислений специальных функций

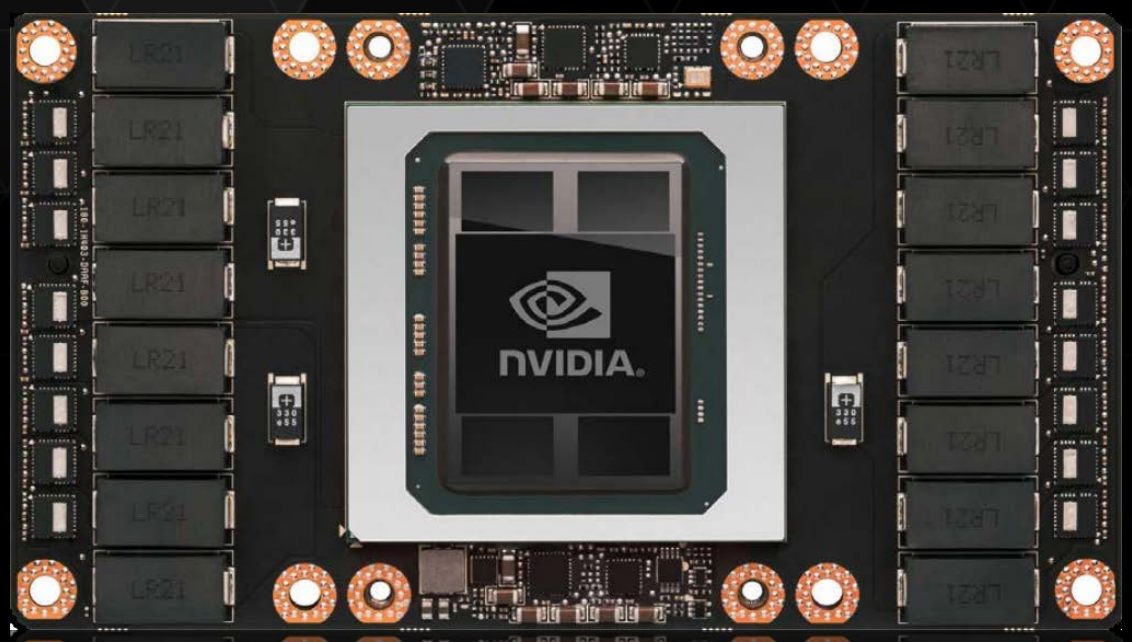
LD/ST - 32 блока загрузки / выгрузки данных

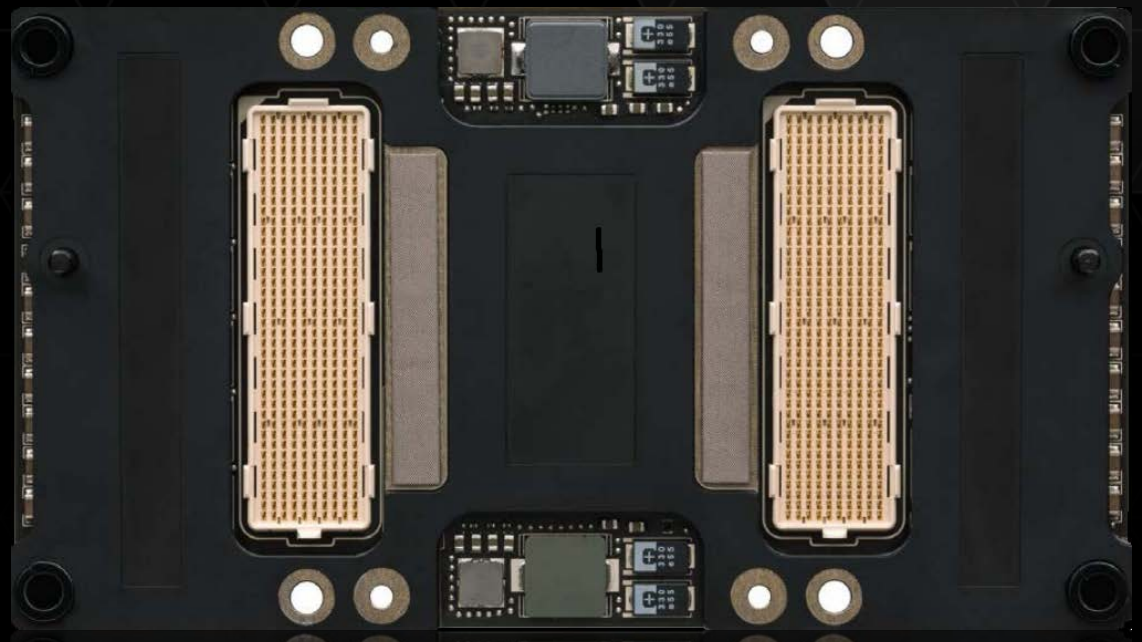
Warp Scheduler - 4 планировщика варпов



ПЛАНИРОВЩИК ВАРПОВ

NVIDIA TESLA P100





АРХИТЕКТУРА PASCAL GP100

Global memory
16 GB

L2 Cache
4 MB

СТРУКТУРА SM GP100

- 3840 CUDA – ядер

Core - 64

DP - 32

SFU - 16

LD/ST - 16

Warp Scheduler - 4

АРХИТЕКТУРА VOLTA GV100

Global memory
16 GB

L2 Cache
6 MB

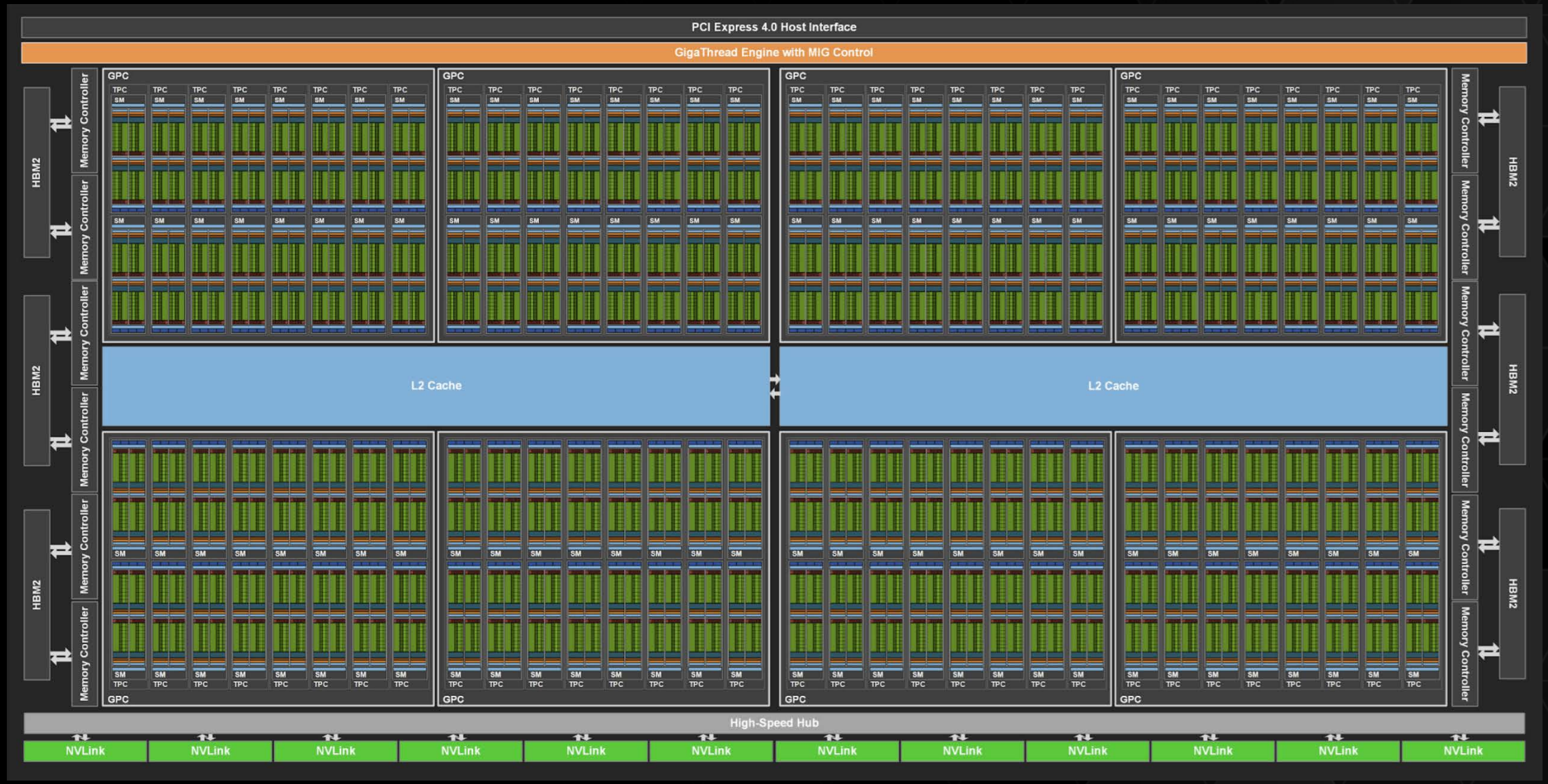
СТРУКТУРА SM GV100

<u>CUDA – ядра</u>

- 5376 FP32
- 5376 INT32
- 2688 FP64
- 672 Tensor cores (FP16/FP32), deep learning

L1 Data Cache / Shared Memory (SM) : 128 КБ
максимальный размер SM : 96 КБ

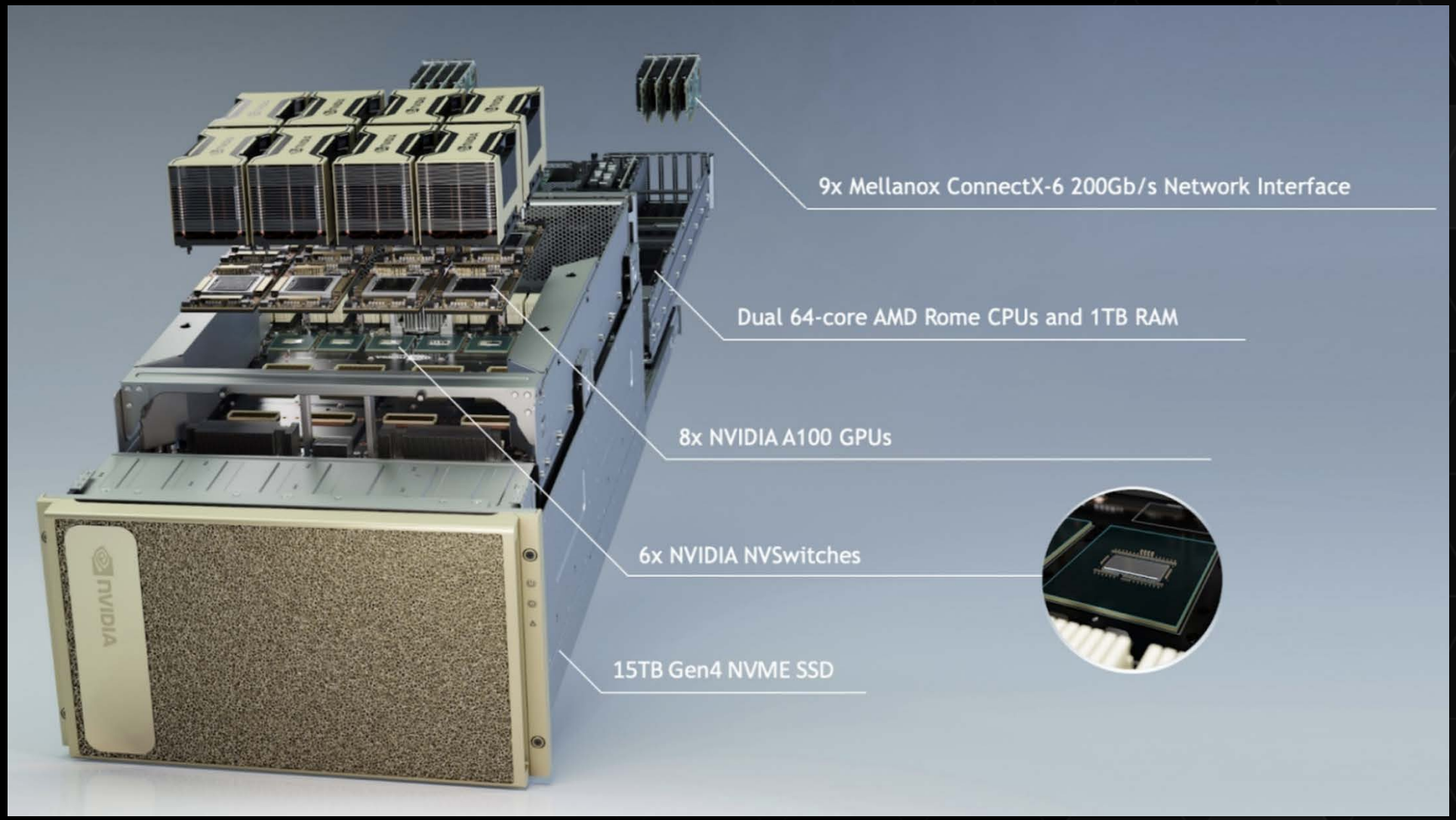
NVIDIA A100


PCI Express 4.0 Host Interface
GigaThread Engine with MIG Control
GPC
TPC
SM
L2 Cache
L2 Cache
HBM2
Memory Controller
High-Speed Hub
NVLink

GA 100

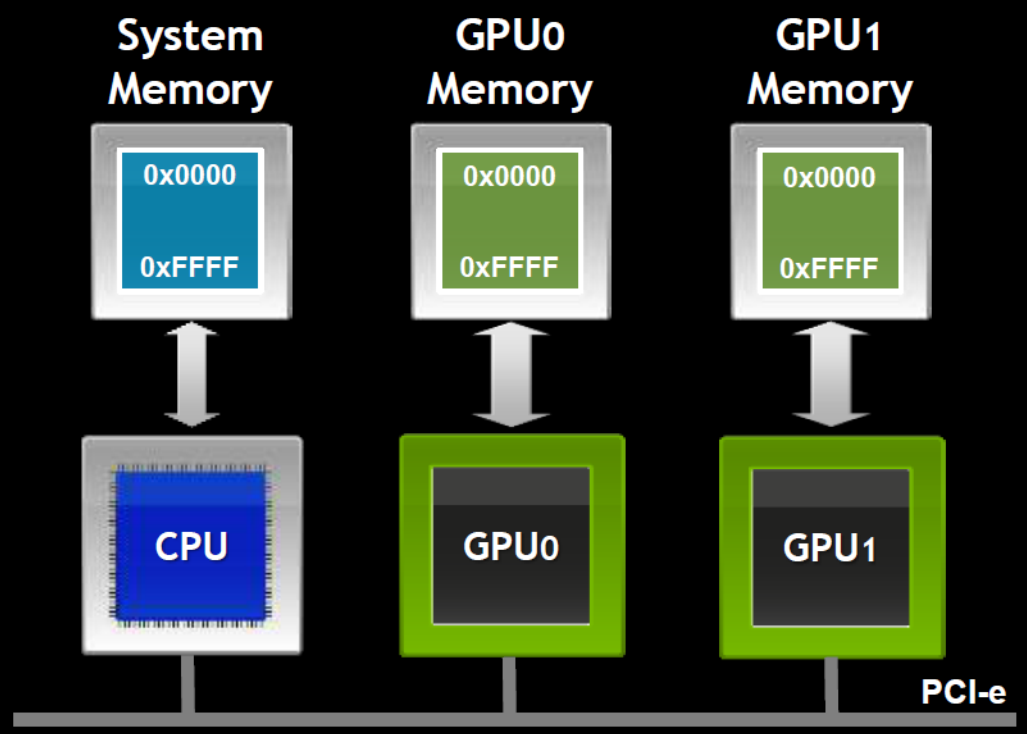
- 8192 FP32
- 8192 INT32
- 4096 FP64

NVIDIA DGX A100 5 PFLOPS

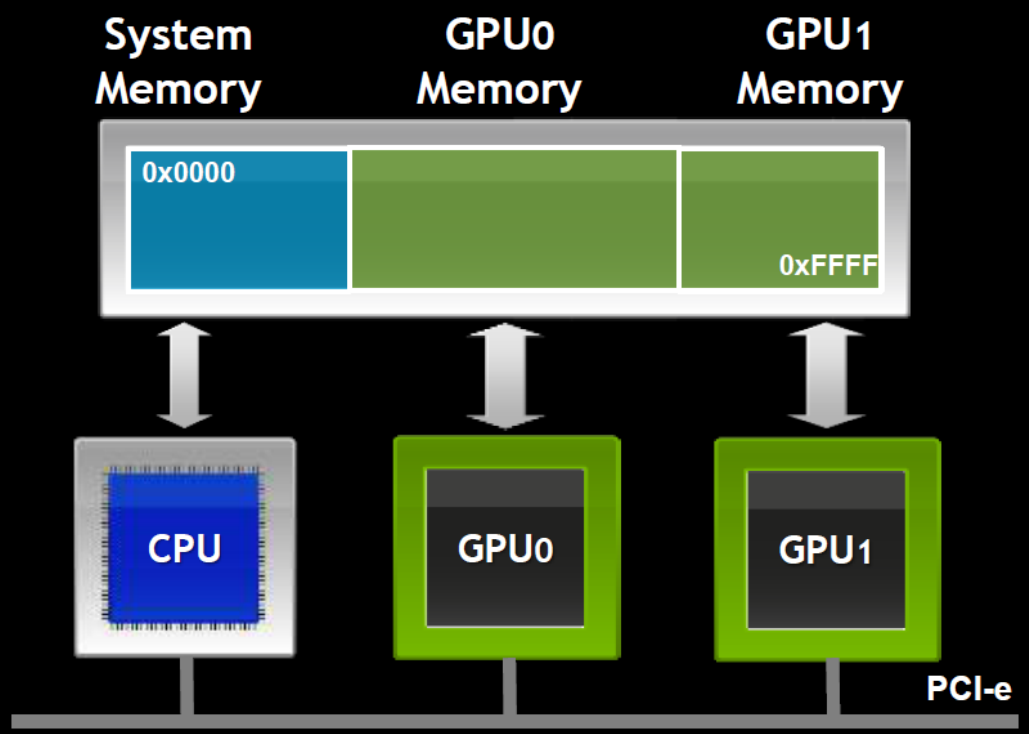


UNIFIED VIRTUAL ADDRESSING

Без UVA

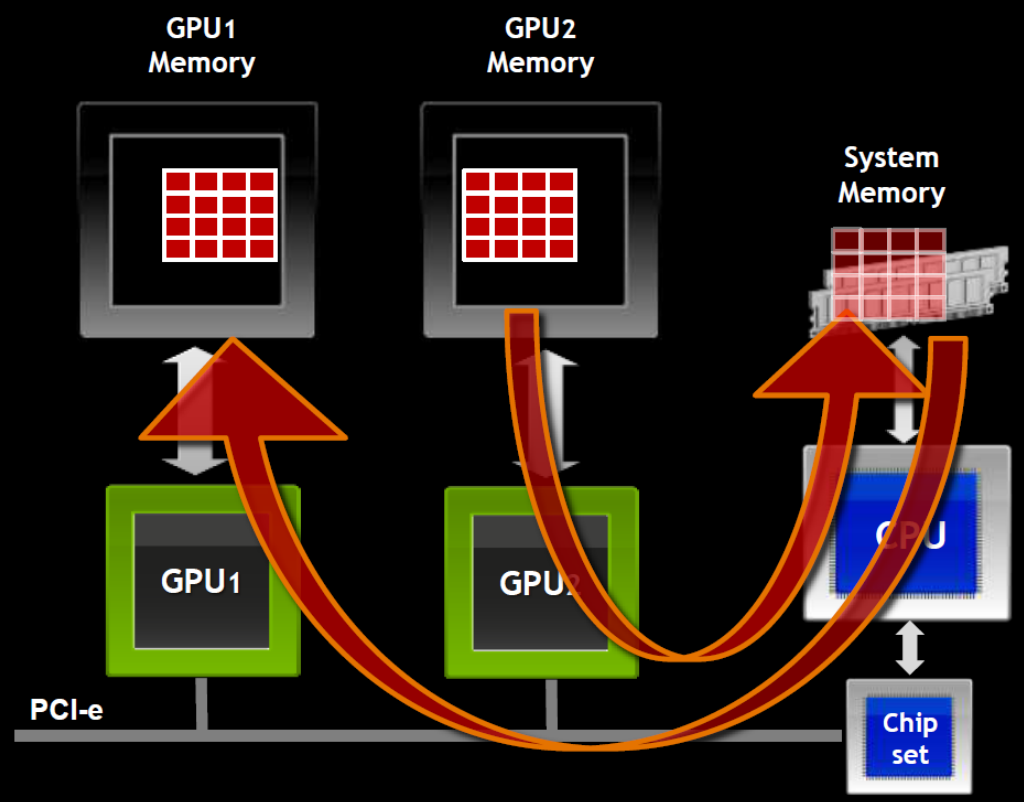


UVA

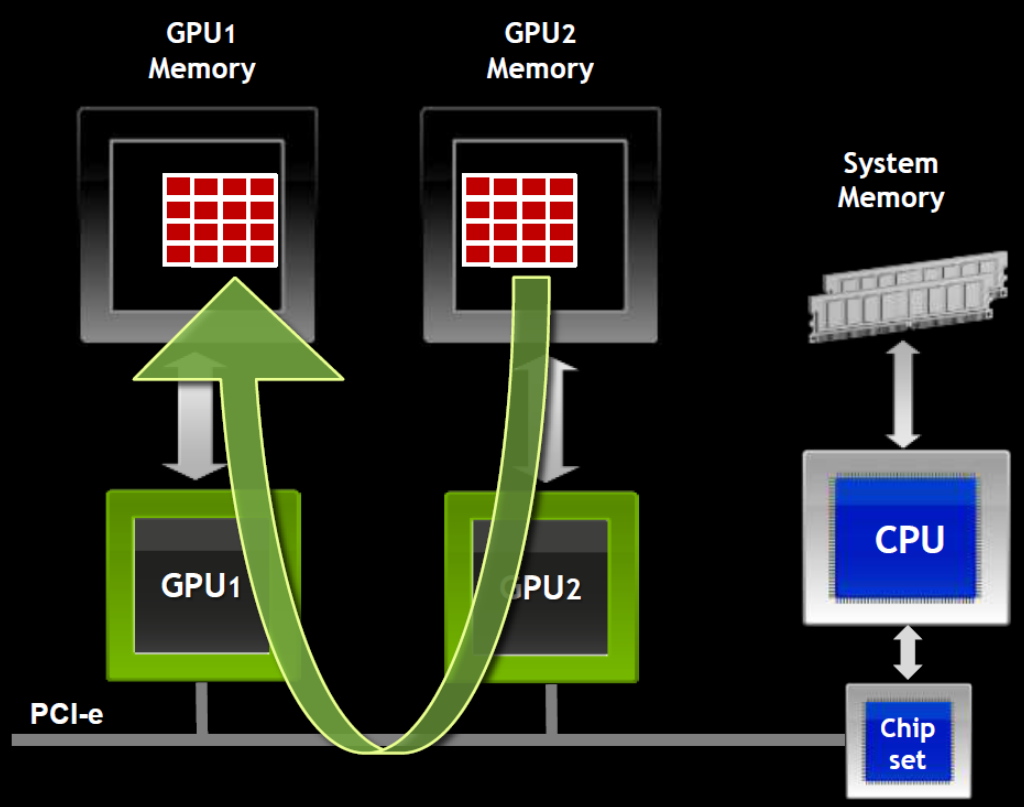


КОПИРОВАНИЕ ДАННЫХ

GPUDirect 1.0

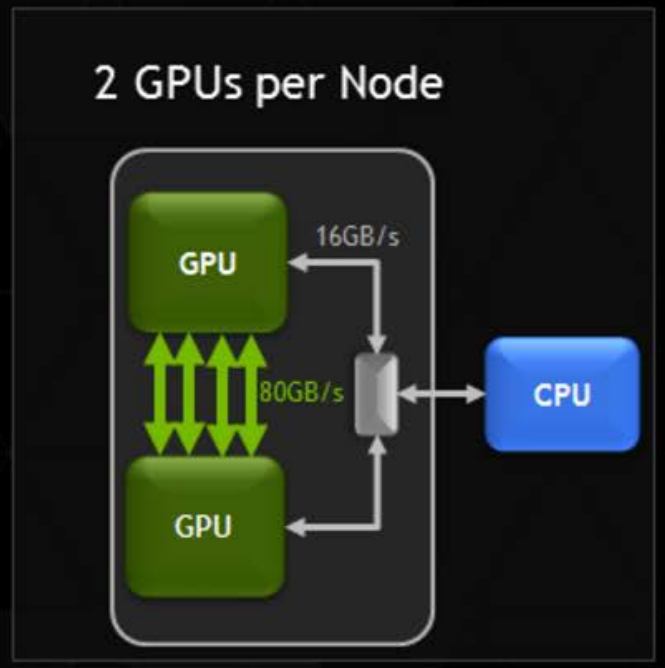


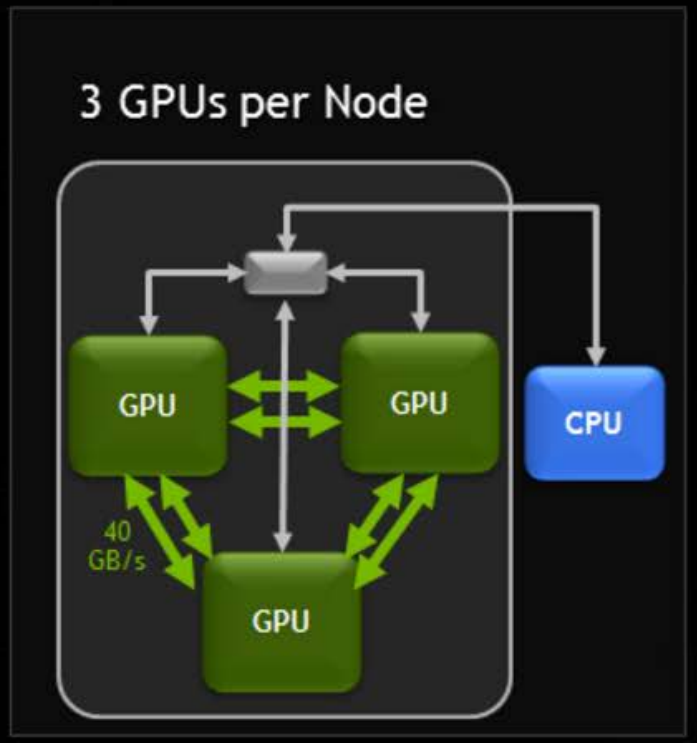
GPUDirect 2.0

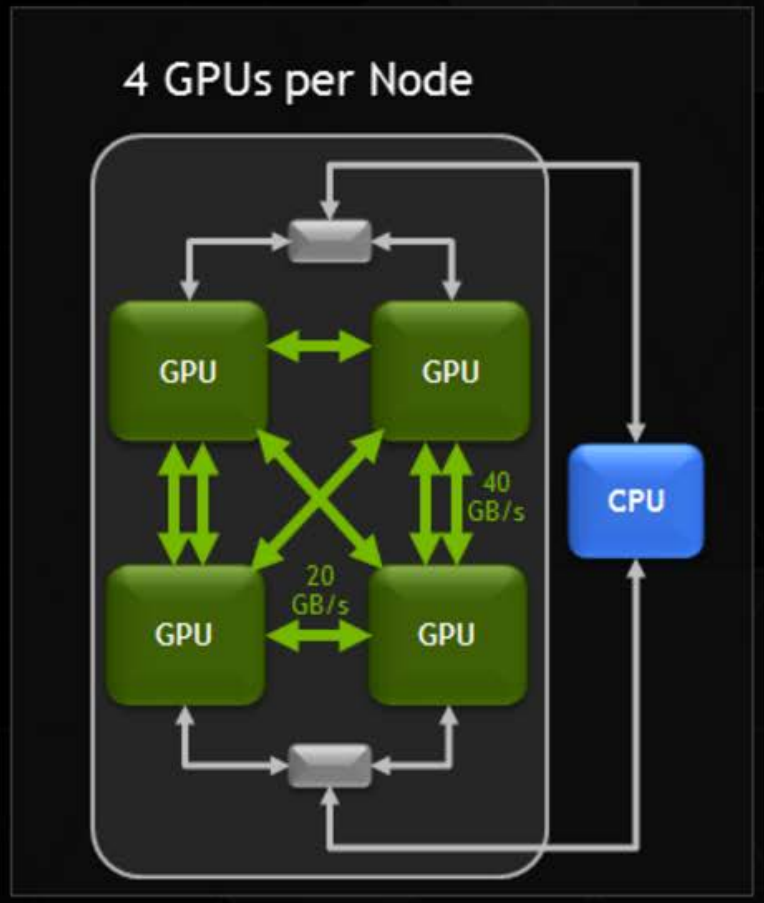


КОПИРОВАНИЕ ДАННЫХ

NVLink







20GB/s NVLink

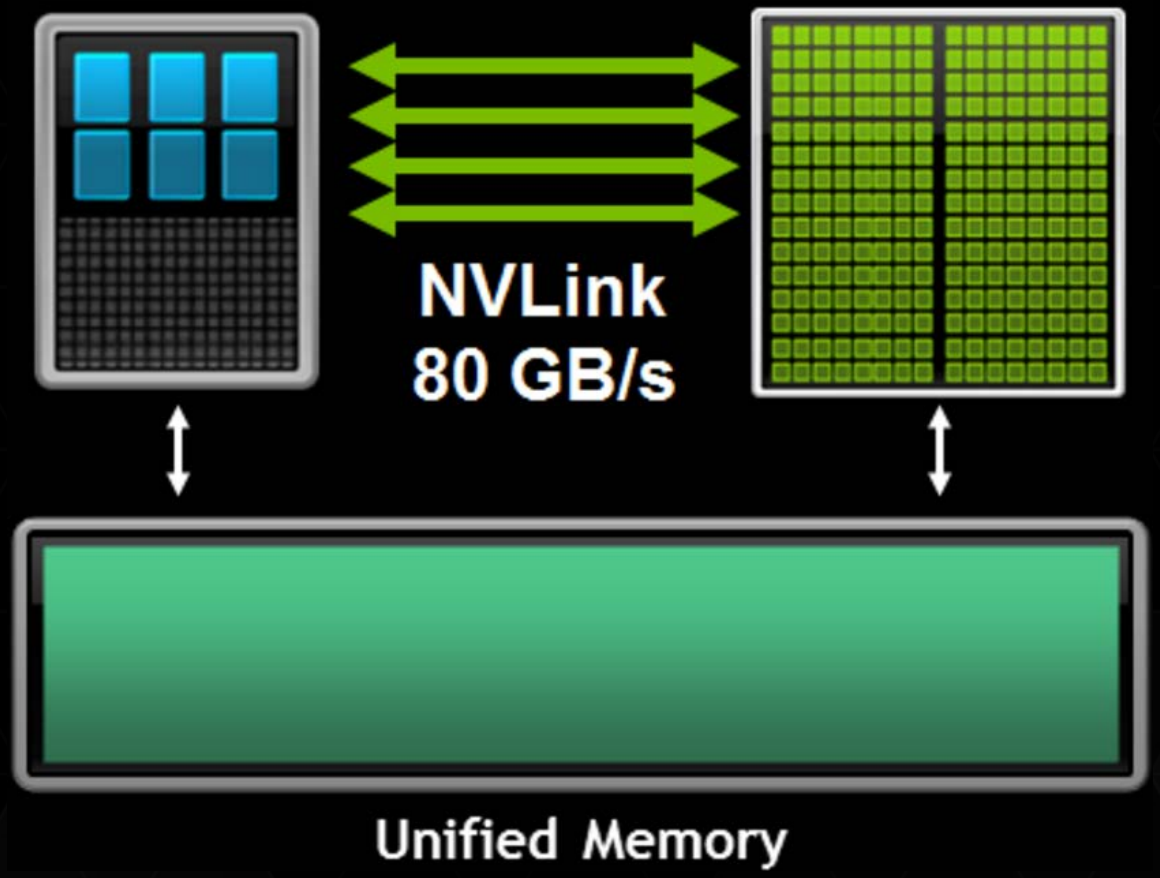
16GB/s PCIe Gen3 x16

PCIe Switch

NVIDIA

КОПИРОВАНИЕ ДАННЫХ

NVLink



Compute Capability	
7.0	TESLA Volta GV100
6.0	TESLA Pascal GP100
5.2	Maxwell GM200
3.0/3.5	TESLA K10/20/20X/40X
2.0/2.1	TESLA Fermi C2070/2090
1.3	TESLA C1060
1.1	Geforce 8800GT
1.0	TESLA C870

Сравнение

	Fermi GF100	Fermi GF104	Kepler GK104	Kepler GK110	Maxwell GM200	Pascal GP100	Volta GV100
Compute Capability	2.0	2.1	3.0	3.5	5.2	6.0	7.0
Threads/Warp	32	32	32	32	32	32	32
Max Warps/Multiprocessor	48	48	64	64	64	64	64
Max Threads/Multiprocessor	1536	1536	2048	2048	2048	2048	2048
Max Thread Blocks/Multiprocessor	8	8	16	16	32	32	32
32-bit Register/Multiprocessor	32768	32768	65536	65536	65536	65536	65336
Max Register/Thread	63	63	63	255	255	255	255
Max Threads/Thread Block	1024	1024	1024	1024	1024	1024	1024
Shared Memory Size Configurations (KB)	16/48	16/48	16/32/48	16/32/48	96	64	Up to 96