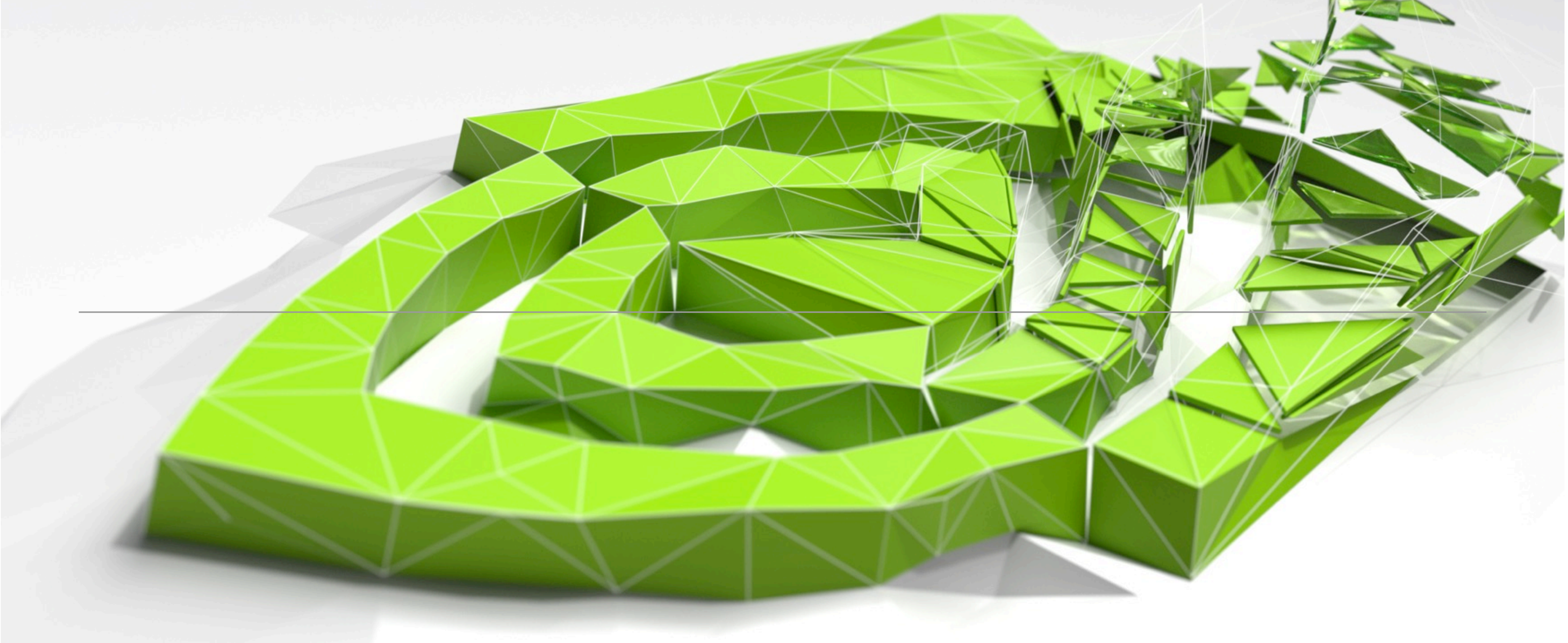
GPUs for HPC, DL and beyond

13th Intl. Conference on Parallel Processing and Applied Mathematics

Bialystok (Poland). September, 8-11, 2019



Manuel Ujaldón

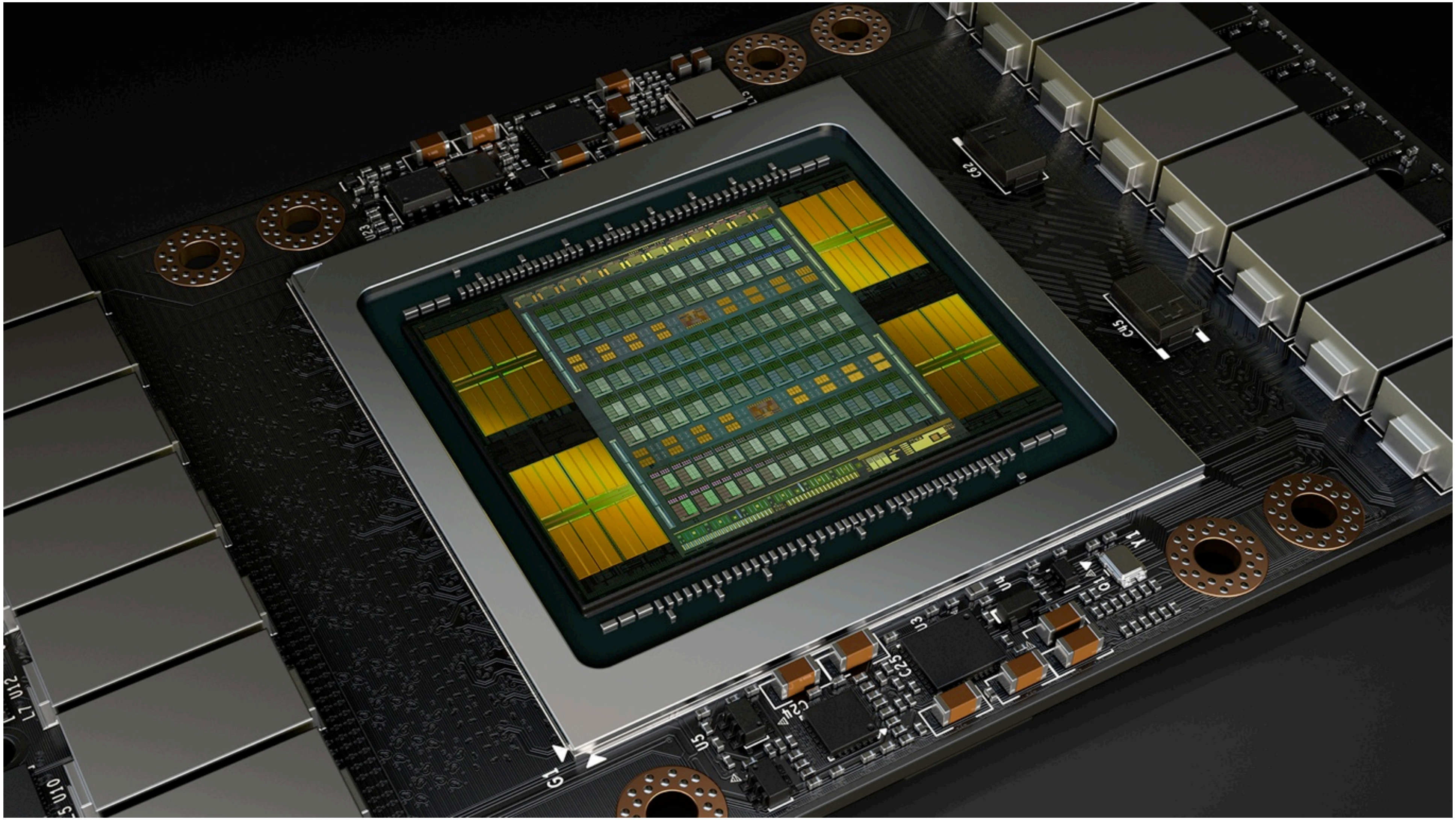
[2016] Full Professor @ Computer Architecture Dept, University of Malaga (Spain)

[2012-2018] CUDA Fellow & [2019] DLI Ambassador @ NVIDIA Corporation (USA)

PPAM 2019

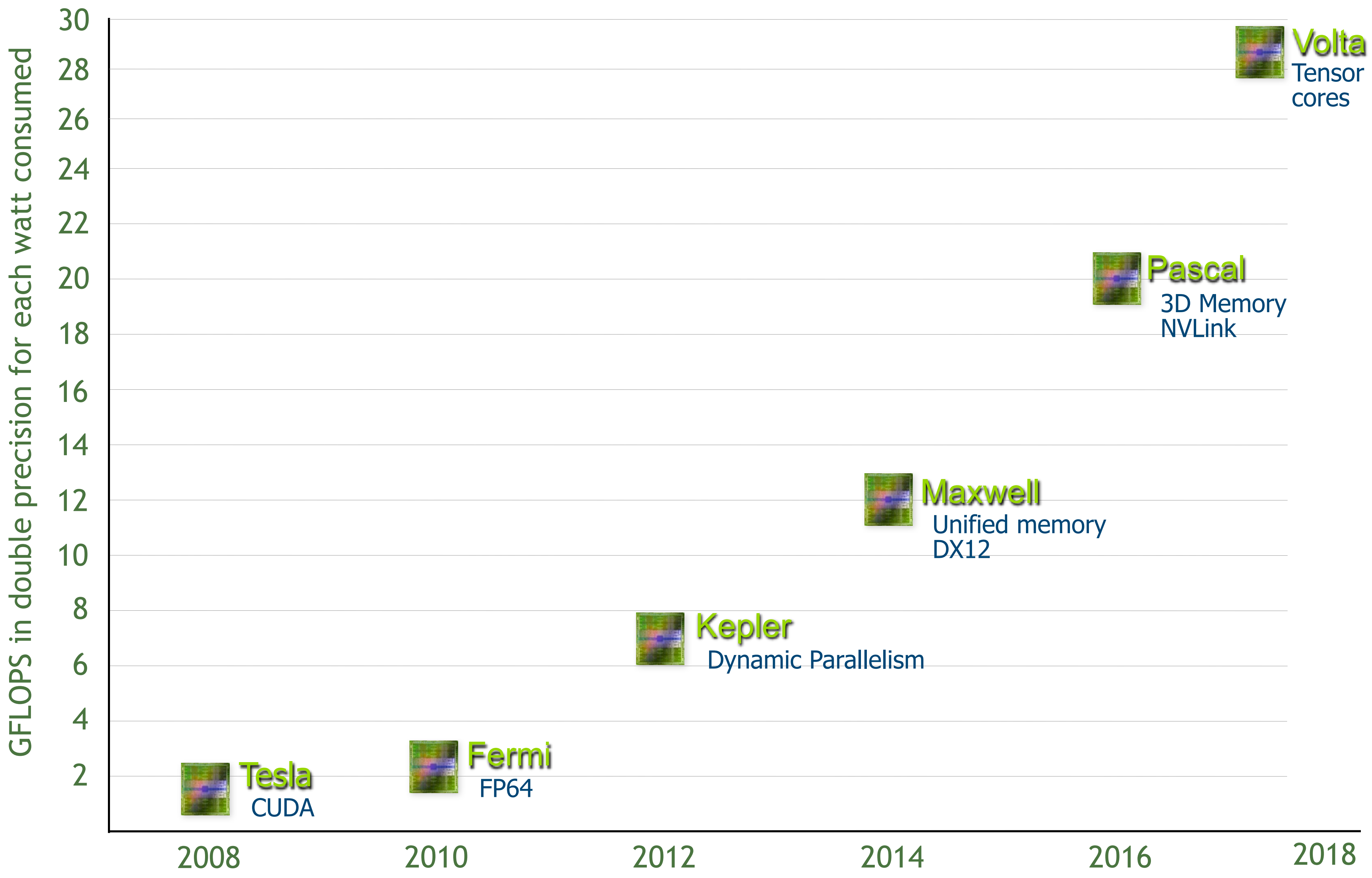
Talk contents [35 slides]

1. Introduction to GPUs [3 slides]
2. The Volta GPU and the new Deep Learning user [14]
3. Turing for everybody [13]
4. Performance analysis based on the roofline model [1]
5. The DGX-1 and DGX-2 supercomputers [2]
6. Summary and conclusions [2].



I. Introduction to GPUs

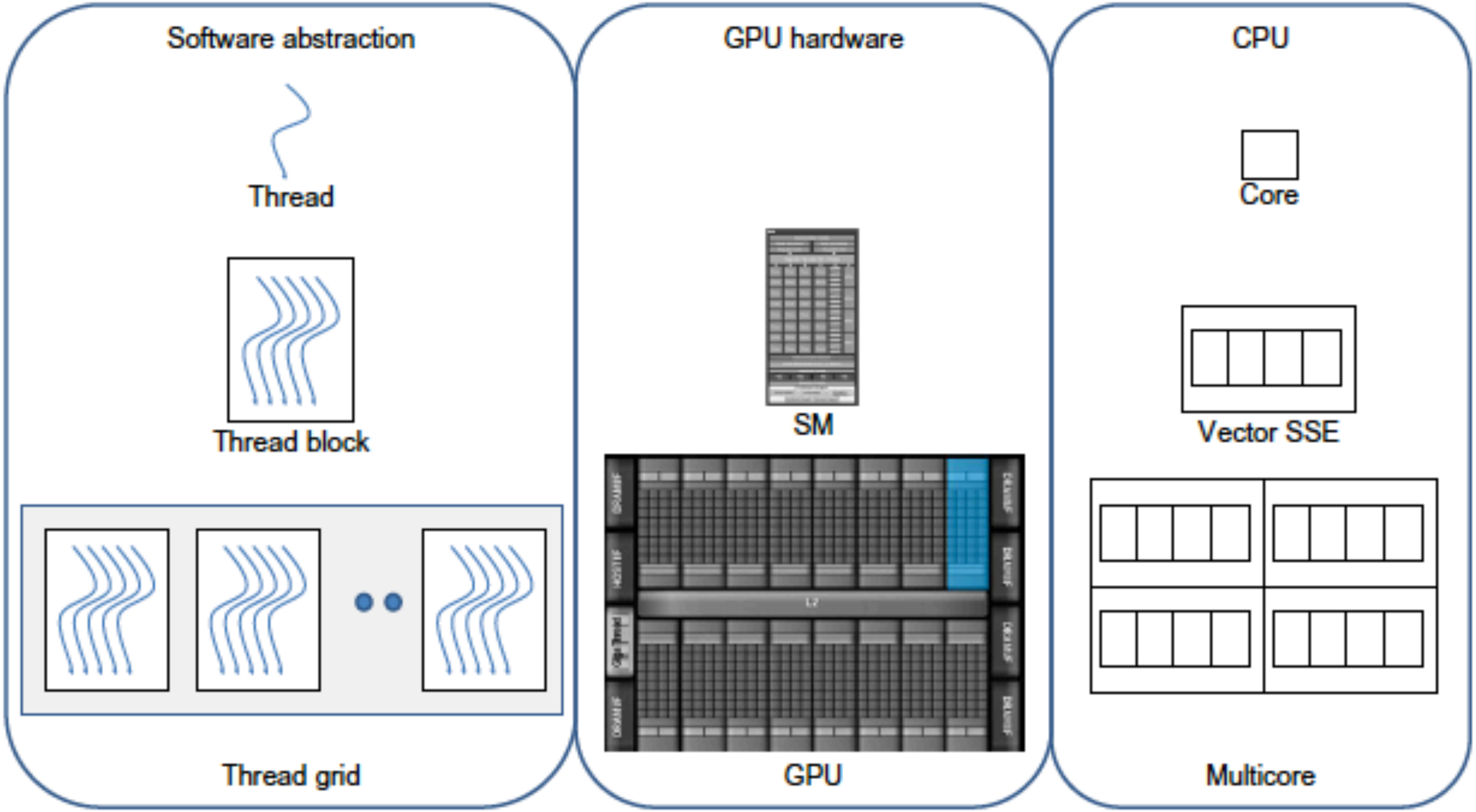
Evolution in performance and energy

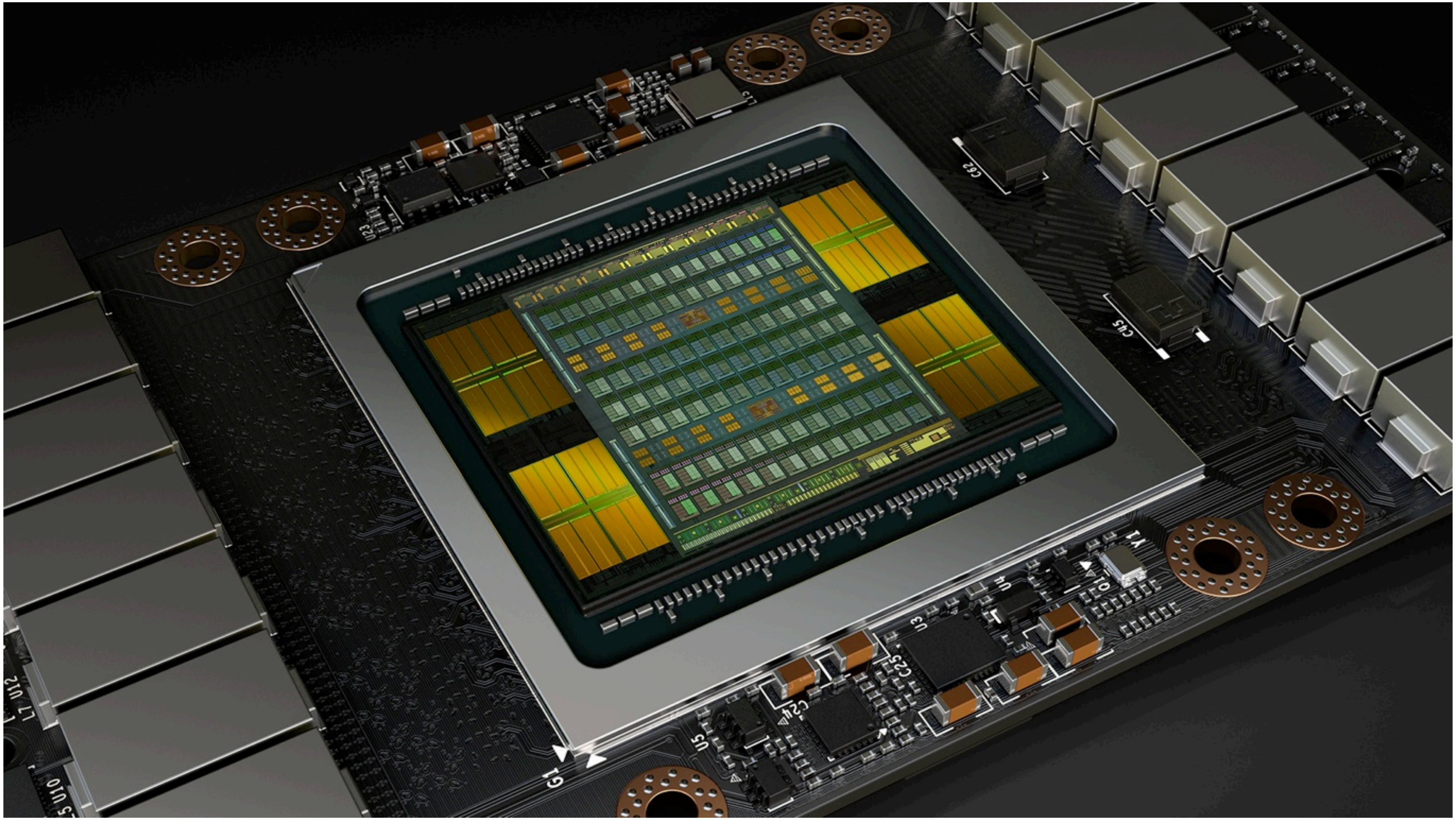


The GPU evolution as a many-core platform

	Maxwell			Pascal			Volta
Architecture	GM107 (GTX750)	GM204 (GTX980)	GM200 (Titan X) (Tesla M40)	GP104 (GTX1080)	GP100 (Titan X) (Tesla P100)	GP102 (Tesla P40)	GV100 (Tesla V100)
Time Frame	2014 /15	2014 /15	2016	2016	2017	2017	2018
CUDA Compute Capability	5.0	5.2	5.3	6.0	6.0	6.1	7.0
N (multiprocs.)	5	16	24	40	56	60	80
M (cores/multip.)	128	128	128	64	64	64	64
Number of cores	640	2048	3072	2560	3584	3840	5120

Comparing the GPU and the CPU: Two methods for building supercomputers





II. The Volta GPU and the Deep Learning user

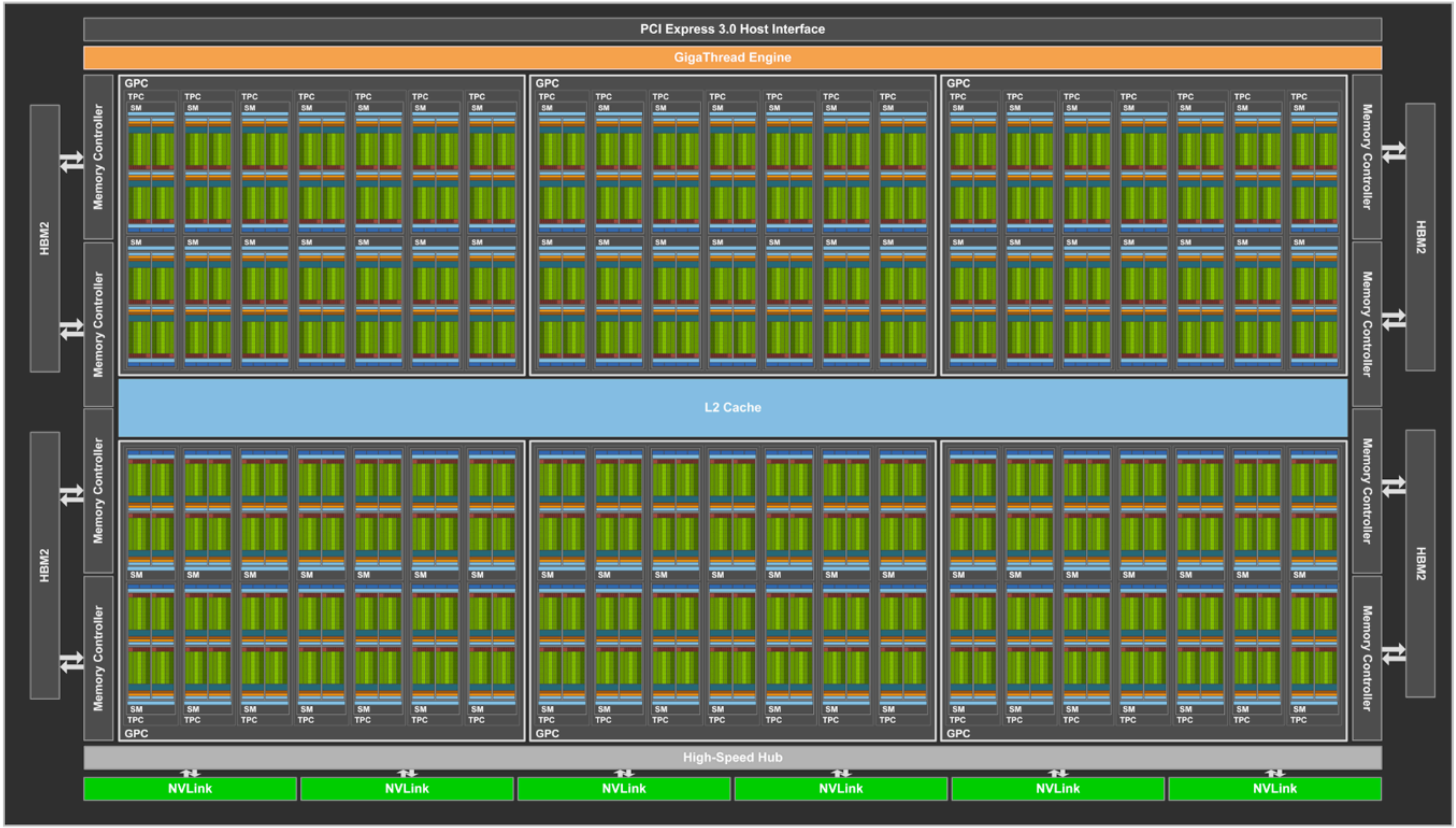
Comparison with Tesla models in previous generations

	K40 (Kepler)	M40 (Maxwell)	P100 (Pascal)	V100 (Volta)
GPU (chip)	GK110	GM200	GP100	GV100
Million of transistors	7100	8000	15300	21100
Die size	551 mm^2	601 mm^2	610 mm^2	815 mm^2
Manufacturing process	28 nm.	28 nm.	16 nm. FinFET	12 nm. FFN
Thermal Design Power	235 W.	250 W.	300 W.	300 W.
Number of fp32 cores	2880 (15 x 192)	3072 (24 x 128)	3584 (56 x 64)	5120 (80 x 64)
Number of fp64 units	960	96	1792	2560
Frequency (regular & boost)	745 & 875 MHz	948 & 1114 MHz	1328 & 1480 MHz	1370 & 1455 MHz
TFLOPS (fp16, fp32, fp64)	No, 5.04, 1.68	No, 6.8, 2.1	21, 10.6, 5.3	30, 15, 7.5
Memory interface	384-bit GDDR5		4096-bit HBM2	
Video memory	Up to 12 GB	Up to 24 GB	16 GB	16 or 32 GB
L2 cache	1536 KB	3072 KB	4096 KB	6144 KB
Shared memory / SM	48 KB	96 KB	64 KB	Up to 96 KB
Register file / SM	65536	65536	65536	65536

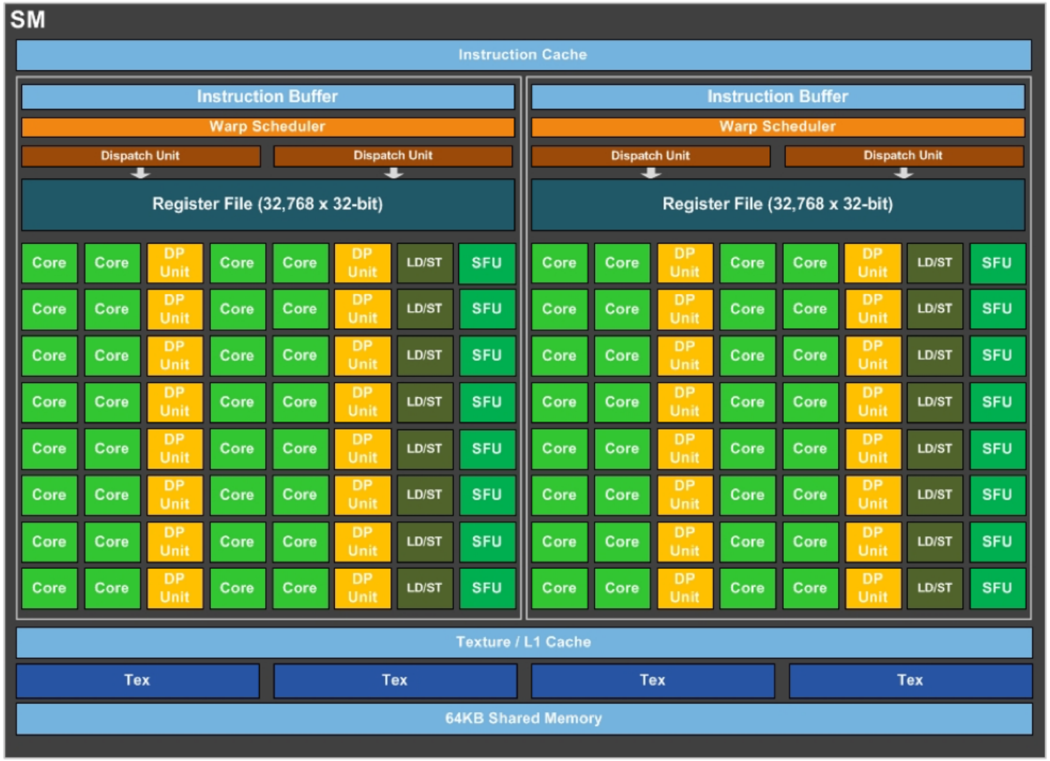
Performance depending on accuracy

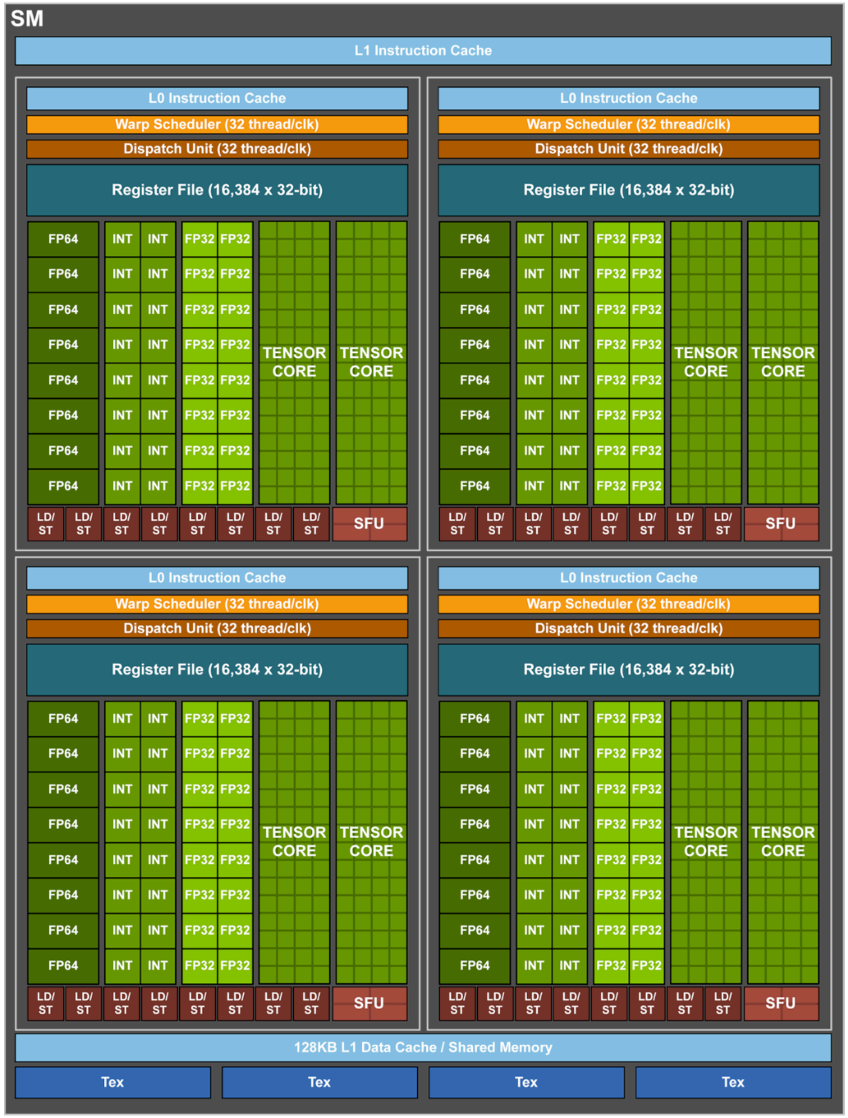
Data type (accuracy)	Tesla P100 (56 SMs)	Tesla V100 (80 SMs)
FP64 (double precision)	32 cores/SM x 1480 MHz x 1 madd = **5.3 TFLOPS**	32 cores/SM x 1455 MHz x 1 madd = **7.5 TFLOPS**
FP32 (single precision)	64 cores/SM x 1480 MHz x 1 madd = **10.6 TFLOPS**	64 cores/SM x 1455 MHz x 1 madd = **15 TFLOPS**
FP16 (half precision)	64 cores/SM x 1480 MHz x 2 madd = **21.2 TFLOPS**	64 cores/SM x 1455 MHz x 2 madd = **30 TFLOPS**
FP16 & 32 in Tensor cores (mixed precision)	None	8 tensor cores/SM x 1455 MHz x 64 madds = **120 TFLOPS**

The GV100 GPU: 84 multiprocessors (SMs) and 8 512-bit memory controllers (Tesla V100 uses only 80 SMs)



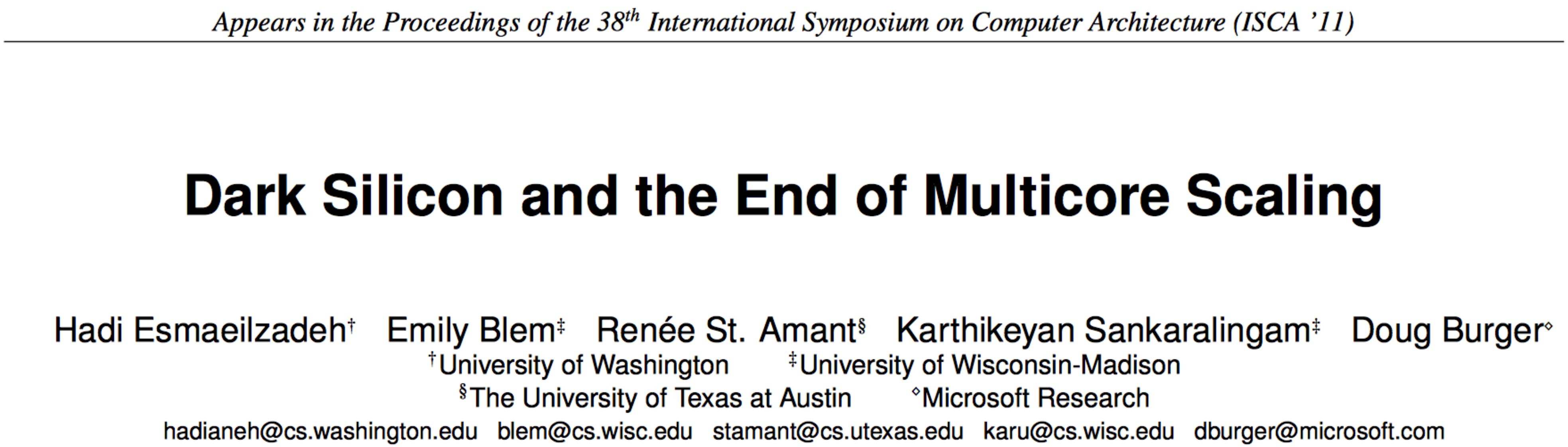
Multiprocessor evolution: From Pascal to Volta




SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
FP64
INT
FP32
TENSOR CORE
LD/ST
SFU
128KB L1 Data Cache / Shared Memory
Tex
Tex
Tex
Tex

The Volta SM partitioning versus Pascal SM

	GP100 SM	GV100 SM
Processing sets ("cloned templates")	2	4
int32 cores / set	32	16
fp32 cores / set	32	16
fp64 cores / set	16	8
Tensor cores / set	None	2
L0 instruction cache / set	None (instruction buffer instead)	1
Register file / set	128 K	64 K
Warp schedulers / set	1	1
Dispatch units / set	1	1

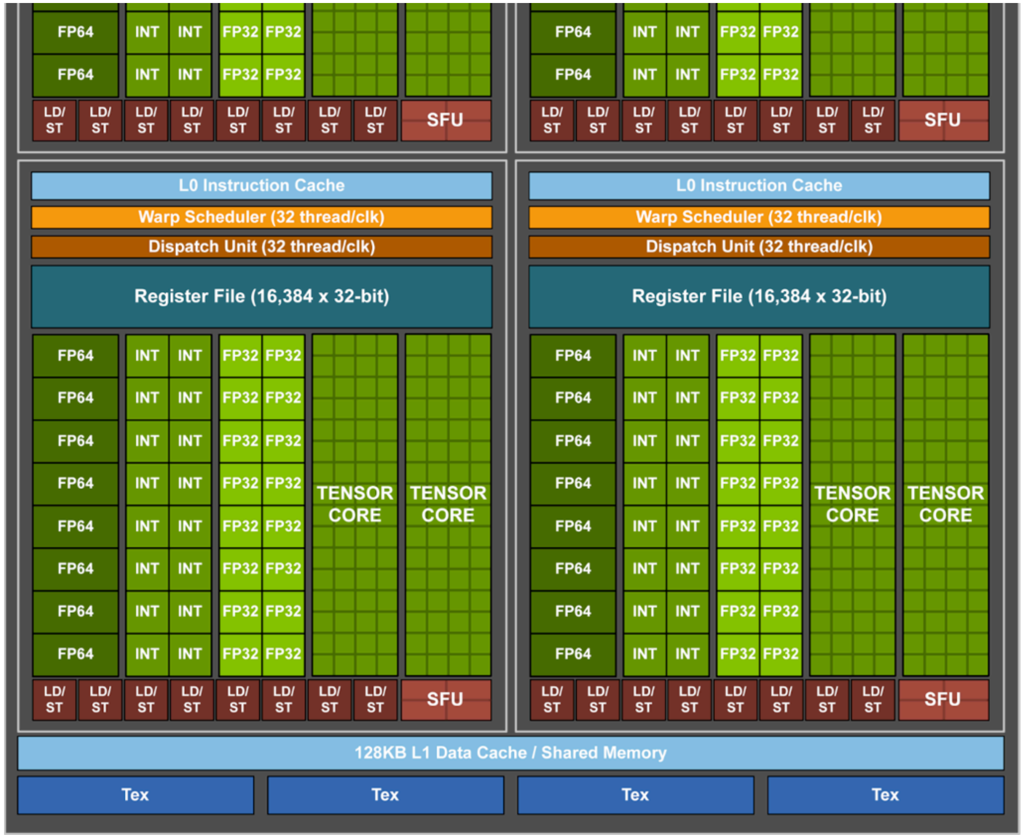


Appears in the Proceedings of the 38th International Symposium on Computer Architecture (ISCA '11)

Dark Silicon and the End of Multicore Scaling

Hadi Esmaeilzadeh† Emily Blem‡ Renée St. Amant§ Karthikeyan Sankaralingam‡ Doug Burger⋄
†University of Washington ‡University of Wisconsin-Madison
§The University of Texas at Austin ⋄Microsoft Research
hadianeh@cs.washington.edu blem@cs.wisc.edu stamant@cs.utexas.edu karu@cs.wisc.edu dburger@microsoft.com

  - 32 fp64 ("double").
  - 8 tensor units.
- We may not use all cores (and in fact we can't). See:
- "Dark silicon at the end of multicore scaling" (ISCA'11)
  - 21% off @ 22 nm. scale.
  - 50% off @ 8 nm. scale.

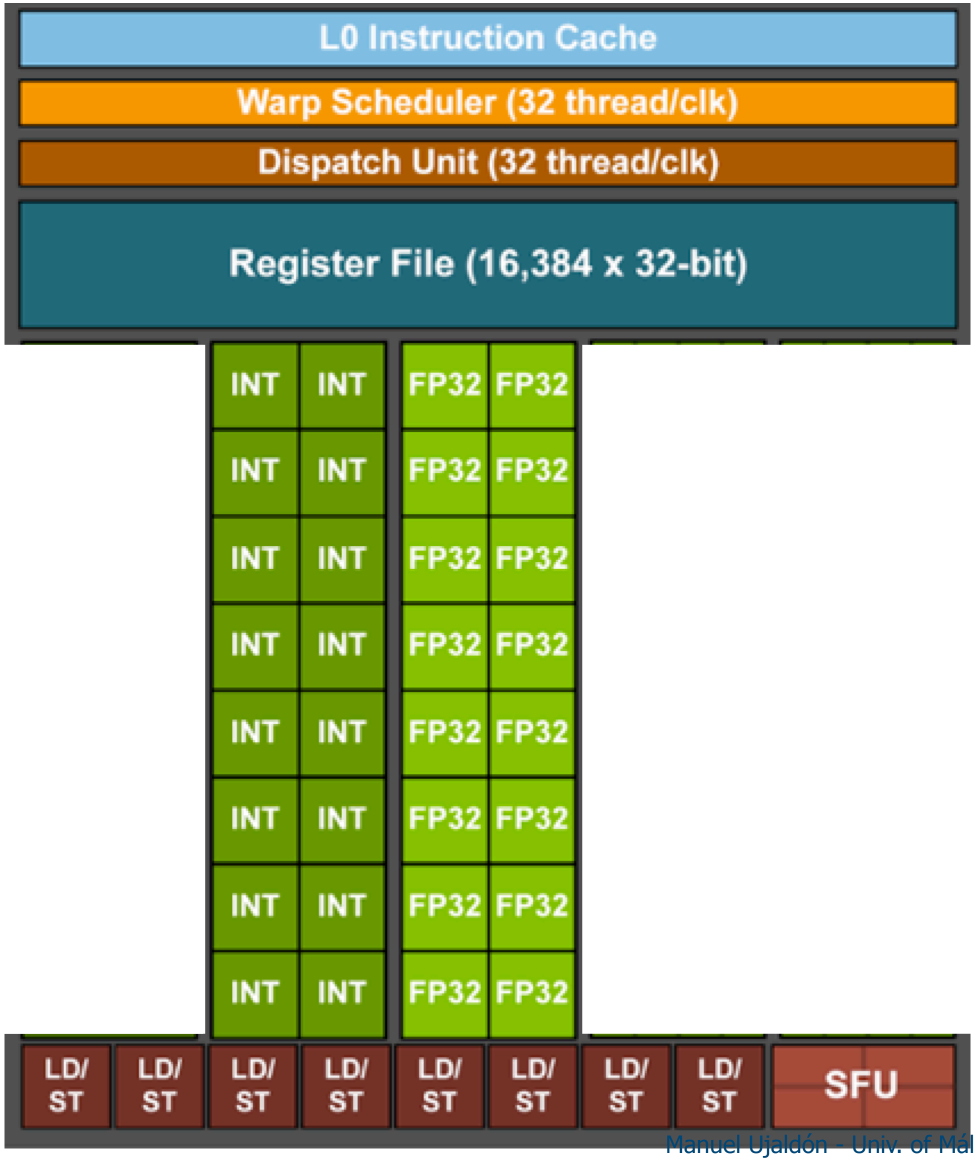

FP64
INT
FP32
LD/ST
SFU
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
TENSOR CORE
128KB L1 Data Cache / Shared Memory
Tex

Volta is a single GPU design for 3 different user profiles: Gamers, HPC and DL

Gamers

HPC scientists

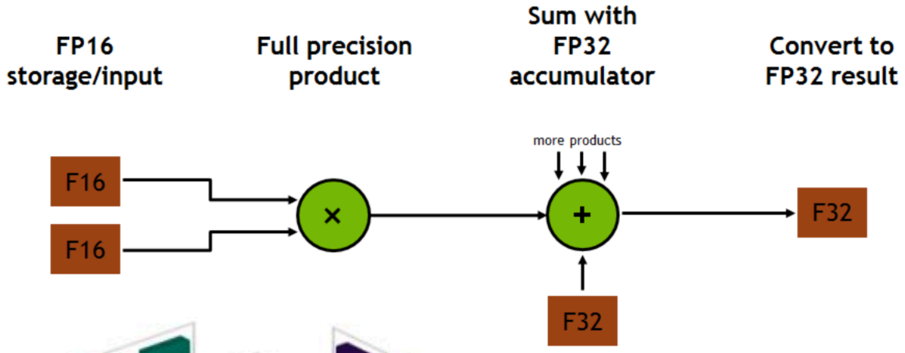
DL users

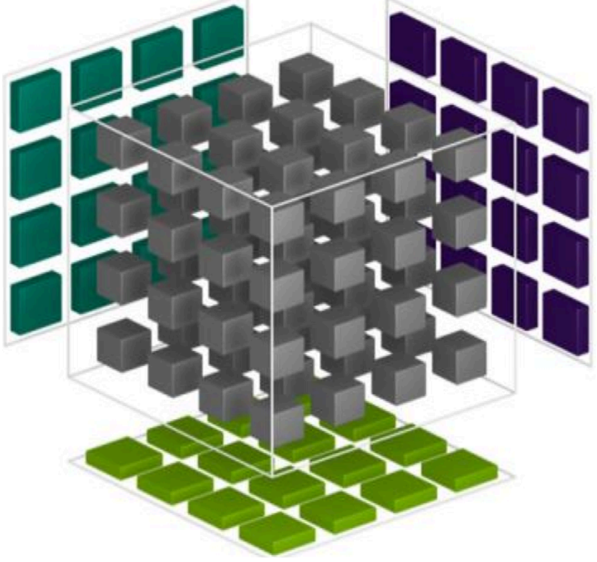

L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
INT
INT
FP32
FP32
LD/ ST
SFU

Tensor cores (8 per multiprocessor in Volta)

- Tensor cores operate on tensors stored in FP16 while computing with FP32, maximizing throughput while keeping the required precision.
- 64 madd mixed-precision ops. per clock (gray cube):
- FP16 input multiply.
- FP32 accumulate.

$$\underbrace{D}_{\text{FP16 or FP32}} = \underbrace{\begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix}}_{\text{FP16}} \underbrace{\begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix}}_{\text{FP16}} + \underbrace{\begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}}_{\text{FP16 or FP32}}$$



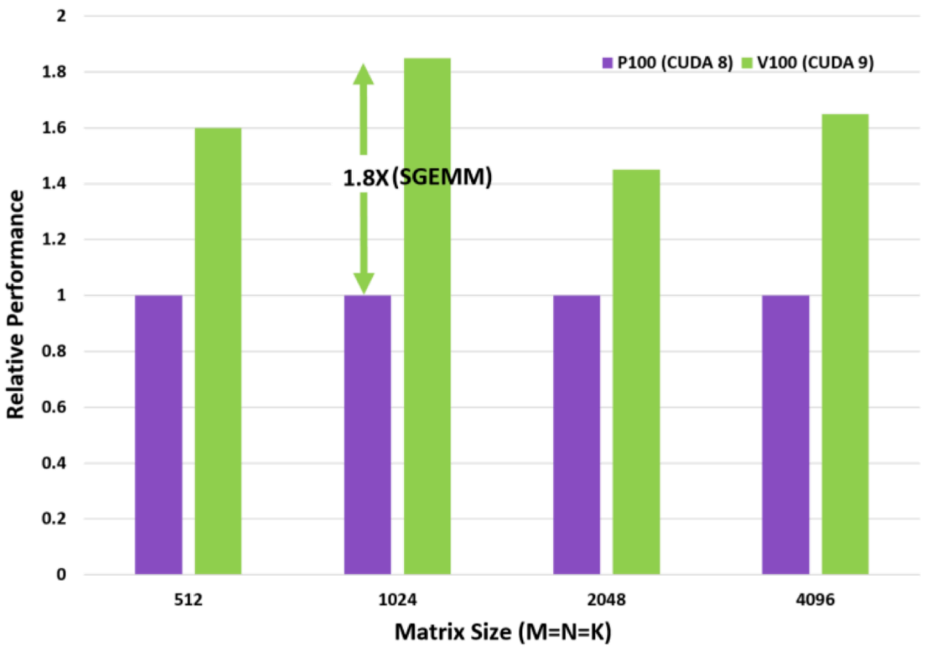


How tensor cores are used

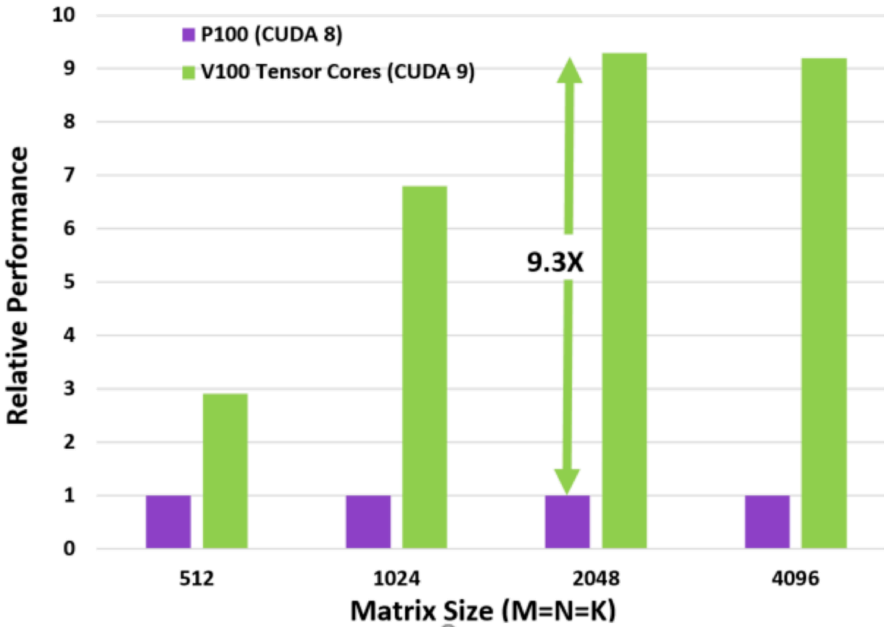
- During program execution, multiple tensor cores are used concurrently by threads within a warp, to compute a larger 16x16x16 matrix operation.
- CUDA exposes these operations as warp-level matrix operations in the CUDA C++ API to provide specialized:
  - Matrix load.
  - Matrix multiply and accumulate.
  - Matrix store.
- Libraries (work at mid-level instead):
  - CUDA 9 cuBLAS and cuDNN extend interfaces to use tensor cores.
- Deep learning frameworks (work even at higher level):
  - Caffe2 and MXNET enable the use of tensor cores on Volta GPUs.

GPU performance on Deep Learning using Tensor cores

- Matrix-Matrix products are extensively used for neural network training and inferencing, to multiply input data and weights in the connected layers of the network.
- Using cuBLAS we can benefit from Volta and tensor cores:
  - FP32:



FP16 input, FP32 compute:



Energy efficiency

- 50% more energy efficient than Pascal.
- New Power Management Modes:
  - Maximum Performance: Operate unconstrained up to its TDP (300W)
  - Maximum Efficiency: Optimal Performance/Watt. A power limit can be set across all GPUs in a rack.

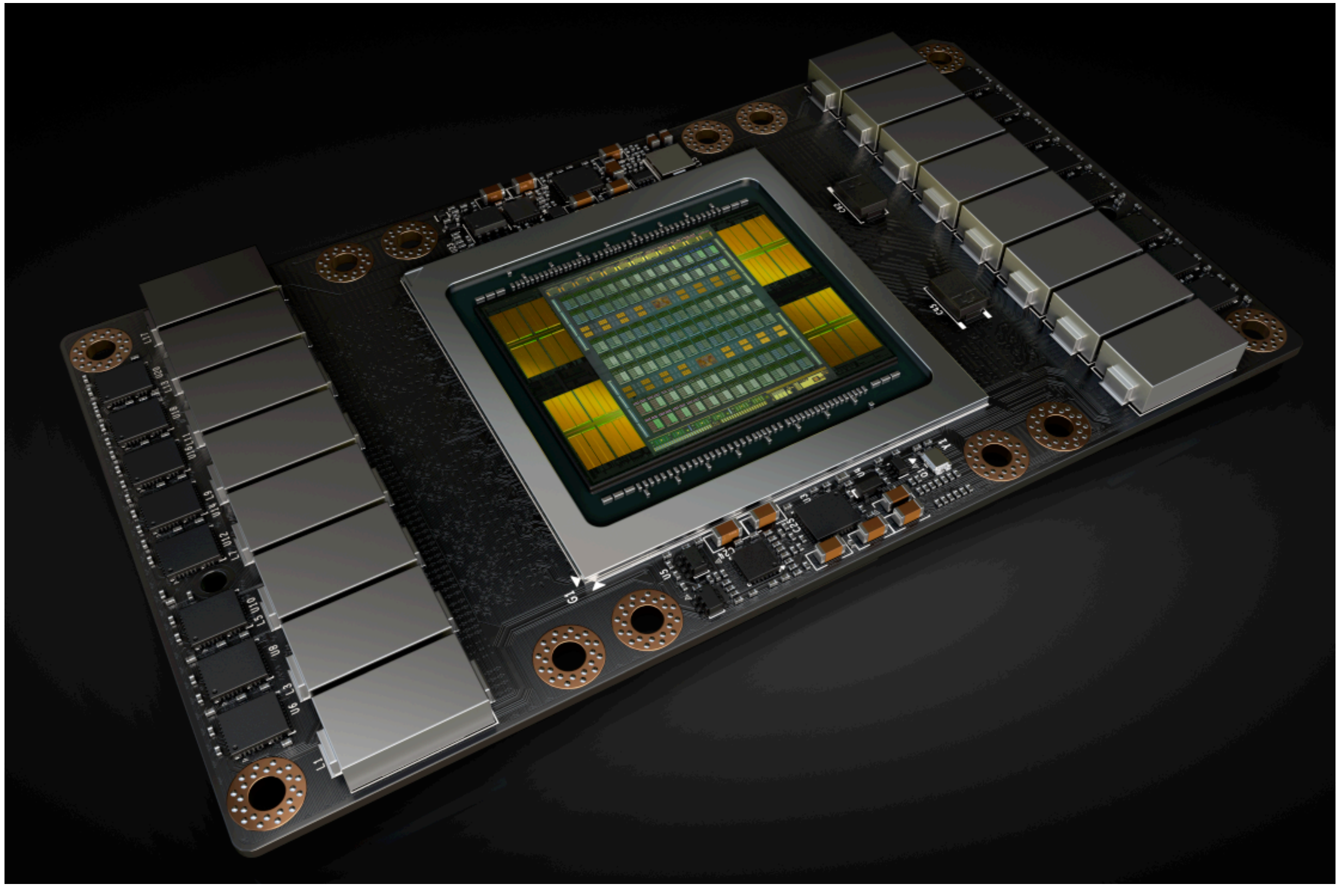
Memory access and performance

- Migration: Unified memory
  - On GV100: New access counters to improve migration of memory pages to the processor accessing most frequently.
  - On IBM Power platforms: New address translation services to allow the GPU to access the CPU's page tables directly.

- Bandwidth: 16 GB HBM2 memory
  - New generation HBM2 memory (from Samsung): 900 GB/s peak bandwidth (1.25x versus 720 GB/s peak in Pascal).
  - New memory controller (from Nvidia): 95%+ bandwidth efficiency running many workloads.

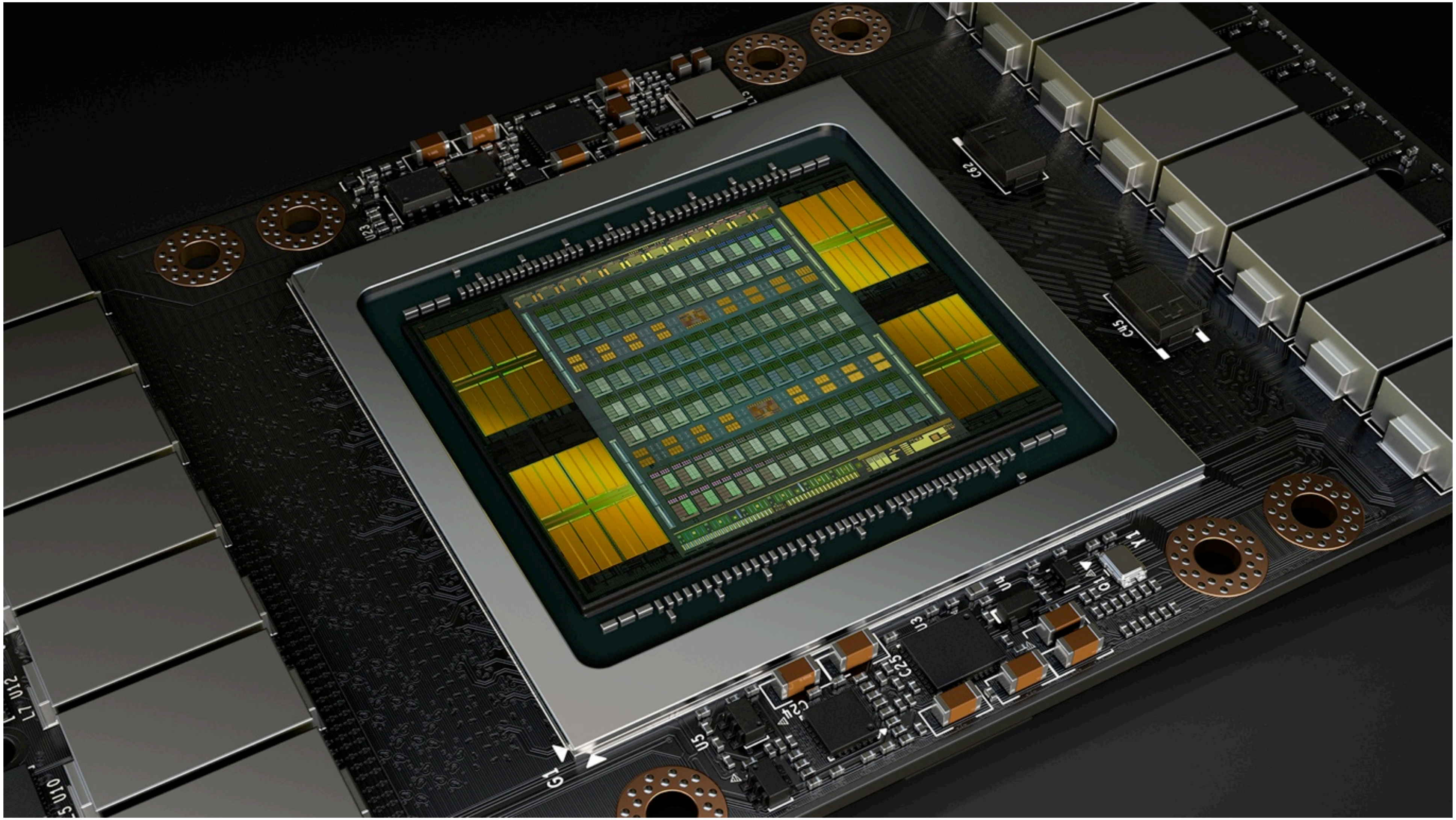
Interconnect: Sockets and slots

- 2nd generation NVLink interconnect with 6 x 25 GB/s. links (vs. 4 x 20 GB/s. in Pascal).



Summary: Volta vs. Pascal

	GP100	GV100	Ratio
FP32 & FP64 peak performance	10 & 5 TFLOPS	15 & 7.5 TFLOPS	1.5x
DL training	10 TFLOPS	120 TFLOPS	12x
DL inferencing	21 TFLOPS	120 TFLOPS	6x
L1 caches (one per multiprocessor)	1.3 MB	10 MB	7.7x
L2 cache	4 MB	6 MB	1.5x
HBM2 bandwidth	720 GB/s	900 GB/s	1.2x
STREAM Triad performance (benchmark)	557 GB/s	855 GB/s	1.5x
NV-link bandwidth	160 GB/s	300 GB/s	1.8x



III. Turing for everybody

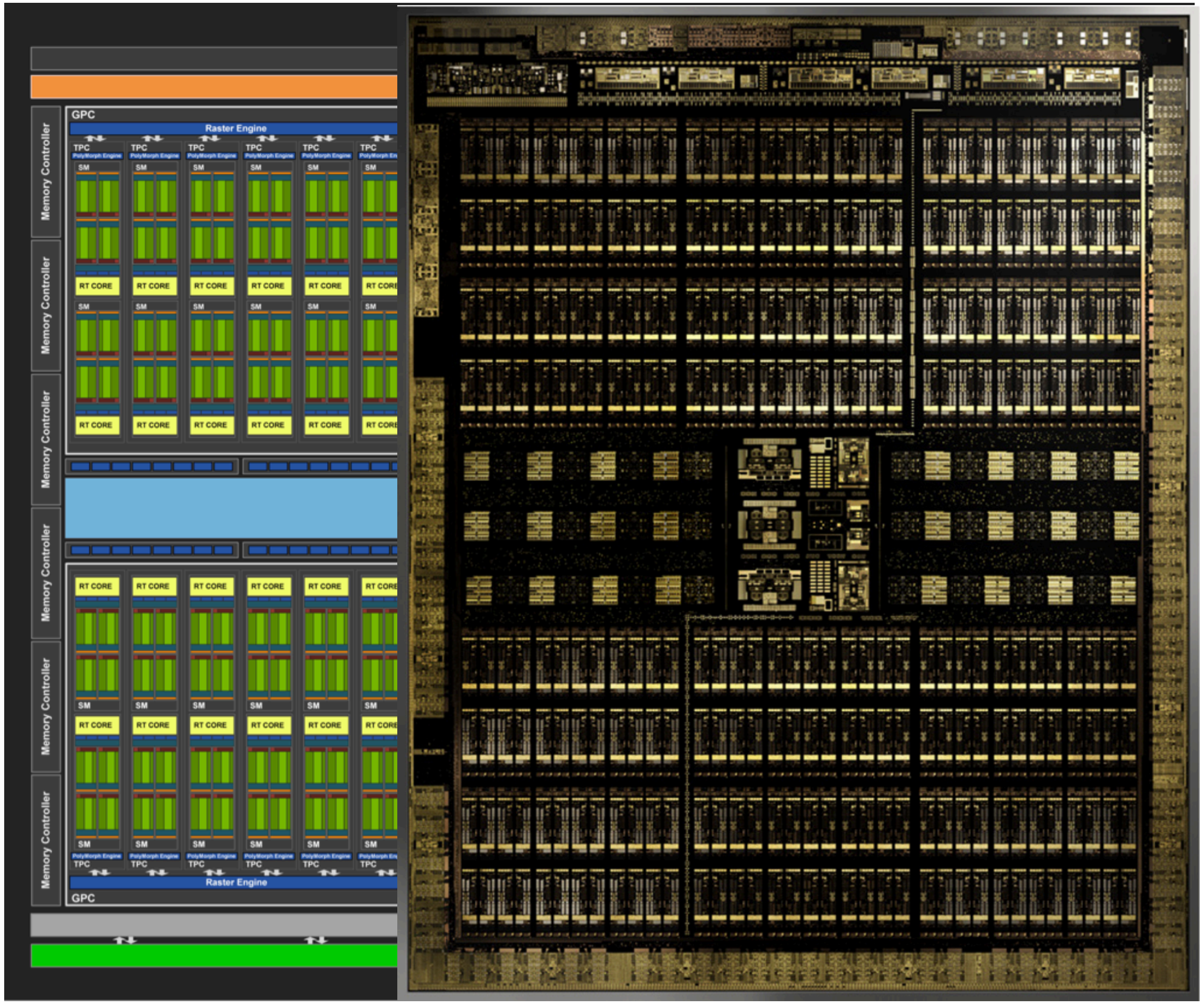
GeForce RTX models

	2060	2060 Super	2070	2070 Super	2080	2080 Super	2080 Ti	Titan RTX
Manufacturing chip	TU-106	TU-106	TU-106	TU-104	TU-104	TU-104	TU-102	TU-102
Release date	Nov'16	Jul'19	Oct'18	Jul'19	Sep'18	Jul'19	Sep'18	Dec'18
Price (USD)	349	399	499	499+	699	699+	999	2499
# multiprocessors (SMs)	30	34	36	40	46	48	68	72
# CUDA cores	1920	2176	2304	2560	2944	3072	4352	4608
GDDR6 memory (GB.)	6	8	8	8	8	8	11	24
Memory bus (bits)	192	256	256	256	256	256	352	384
Mem. bandwidth (GB/s.)	336	448	448	448	448	496	616	672

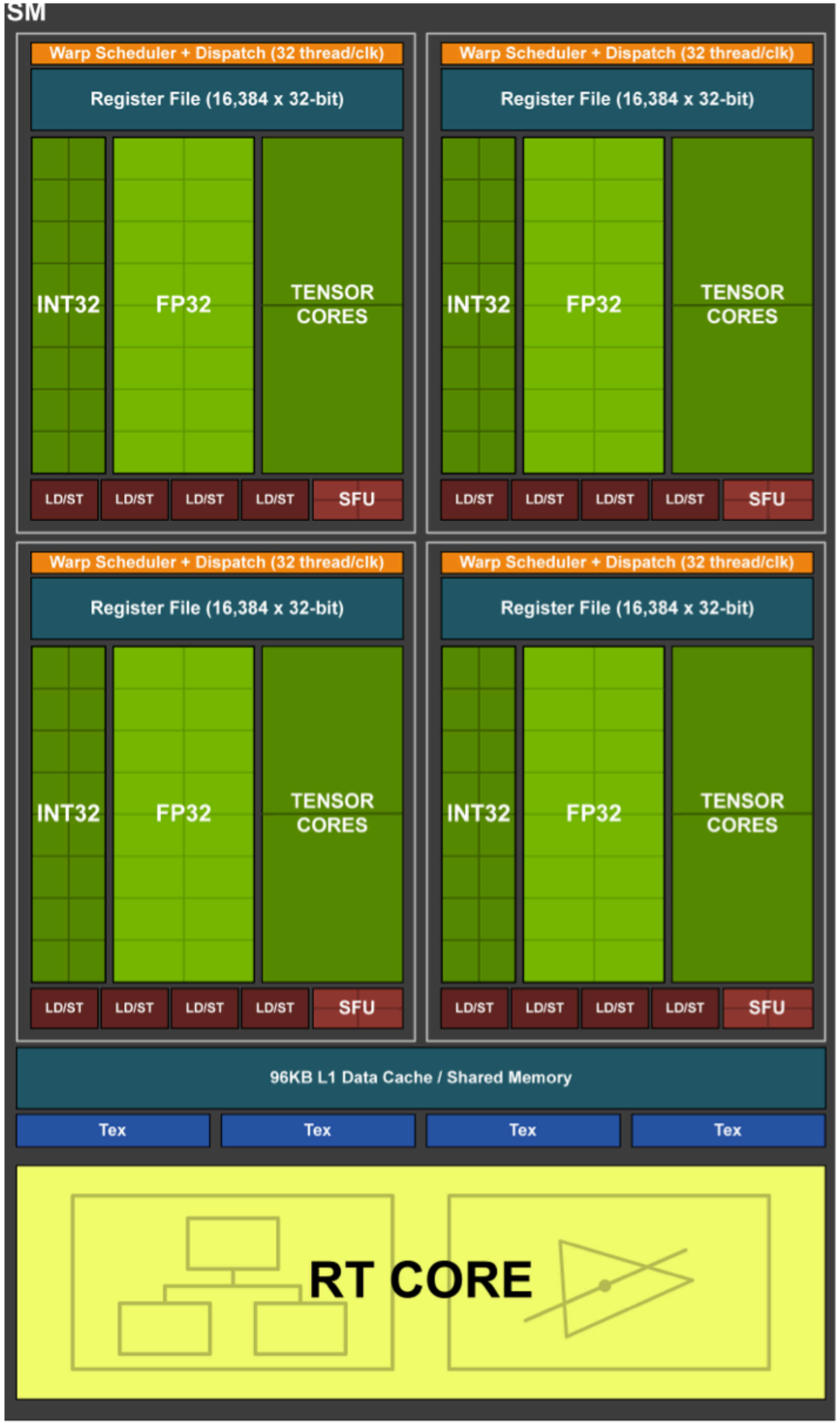
- Our analysis is based on Turing TU102 flagship chip.

The TU-102 chip (72 Turing multiprocessors or SMs)

- 4608 CUDA cores (64 per SM).
- 576 Tensor cores (8 per SM).
- 72 Ray Tracing cores (1 per SM).
- 288 texture units (4 per SM).
- 12 32-bit GDDR6 memory controllers (384 bits total).



The Turing multiprocessor


SM
Warp Scheduler + Dispatch (32 thread/clk)
Register File (16,384 x 32-bit)
INT32
FP32
TENSOR CORES
LD/ST
LD/ST
LD/ST
LD/ST
SFU
Warp Scheduler + Dispatch (32 thread/clk)
Register File (16,384 x 32-bit)
INT32
FP32
TENSOR CORES
LD/ST
LD/ST
LD/ST
LD/ST
SFU
Warp Scheduler + Dispatch (32 thread/clk)
Register File (16,384 x 32-bit)
INT32
FP32
TENSOR CORES
LD/ST
LD/ST
LD/ST
LD/ST
SFU
Warp Scheduler + Dispatch (32 thread/clk)
Register File (16,384 x 32-bit)
INT32
FP32
TENSOR CORES
LD/ST
LD/ST
LD/ST
LD/ST
SFU
96KB L1 Data Cache / Shared Memory
Tex
Tex
Tex
Tex
RT CORE

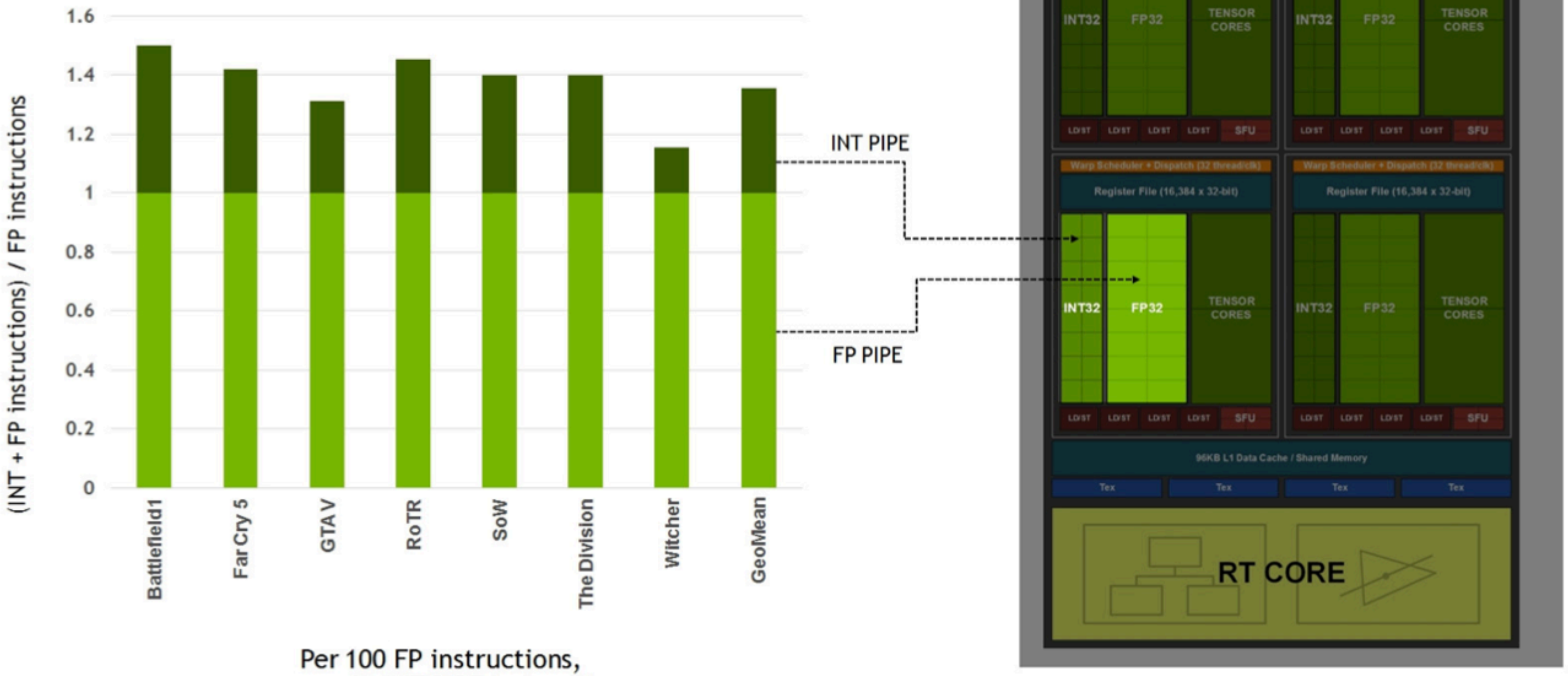
Legacy from Volta

- Independent Thread Scheduling.
- Hardware-accelerated Multi-Process Service (MPS) with address space isolation for multiple applications.
- Cooperative Groups.
- NV-link to provide high bandwidth and low latency connectivity between pairs of Turing GPUs (up to 100 GB/s bidirectional bandwidth).

Key architectural changes: Computation

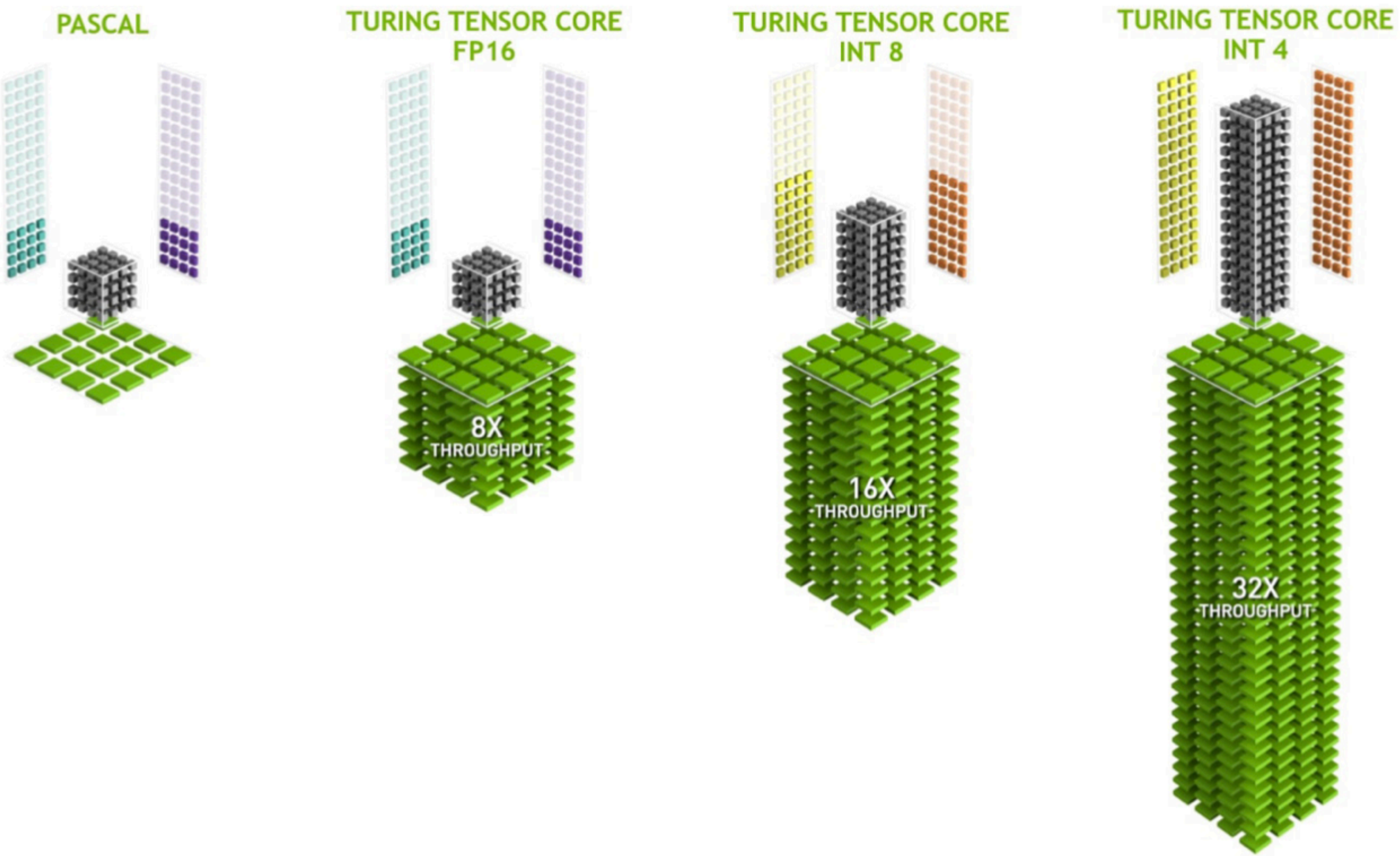
- Independent integer **datapath** that can execute instructions concurrently with the FP datapath. In previous generations, executing ints. blocked FP instrs. from issuing.
- Tensor cores add new **INT8 and INT4 precision** modes for inferencing workloads that can tolerate quantization and do not require FP16 precision.

CONCURRENT EXECUTION



Per 100 FP instructions,
average 36 INT PIPE instructions
(ie iadd, select, fp min/max, compare etc)

INT8 and INT4 precision on Tensor cores



Key architectural changes: Memory

- Caches. Unify shared **memory**, texture caching and memory load caching into a single unit for a 2x bandwidth and 2x capacity available for L1 cache.
- DRAM. First GPU to support **GDDR6** memory:
  - 7 GHz clock rate and DDR means 14 Gbps.
  - 12 memory controllers x 32 wires/m.c. x 14 Gbps/wire = **672 GB/s.**
  - 20% improved power efficiency vs GDDR5X in Pascal.

Major hardware improvements

- Biggest architectural leap forward in over a decade. Major advances for:
  - GeForce users: Efficiency and performance for PC gaming.
  - Quadro users: Professional graphics applications.
  - GPGPU users: Deep learning and HPC acceleration.
- New accelerators and a hybrid rendering approach to fuse:
  - Real-time ray tracing.
  - AI.
  - Rasterization and simulations.
- New GPU multiprocessor:
  - Introducing Ray Tracing cores.
  - Integrating memory units.
  - Improving shader execution.

Major software improvements

- Tensor cores power a suite of Neural Services:
  - Stunning graphics effects for gamers and professionals.
  - Fast AI inferencing for cloud-based systems.
- New Ray Tracing cores for real-time ray-traced rendering, combined with DirectML for AI and DirectX Raytracing (DXR) APIs by Microsoft (from early 2018 on).
- Advanced shading features to:
  - Improve performance.
  - Enhance image quality.
  - Deliver new levels of geometric complexity.

Clarifying the mess of clock frequencies

Commercial product	**Clock type**	**GeForce**		**Quadro**	
		GTX 1080 Ti [Pascal]	**RTX 2080 Ti [Turing]**	**P6000 [Pascal]**	**RTX 6000 (and GeF. Titan RTX) [Turing]**
Reference model	GPU Base	1480	1350	1506	1455
	GPU Boost	1582	1545	1645	1770
Founders Edition	GPU Base	1480	1350	1506	1455
	GPU Boost	1582	**1635**	1645	**1770**

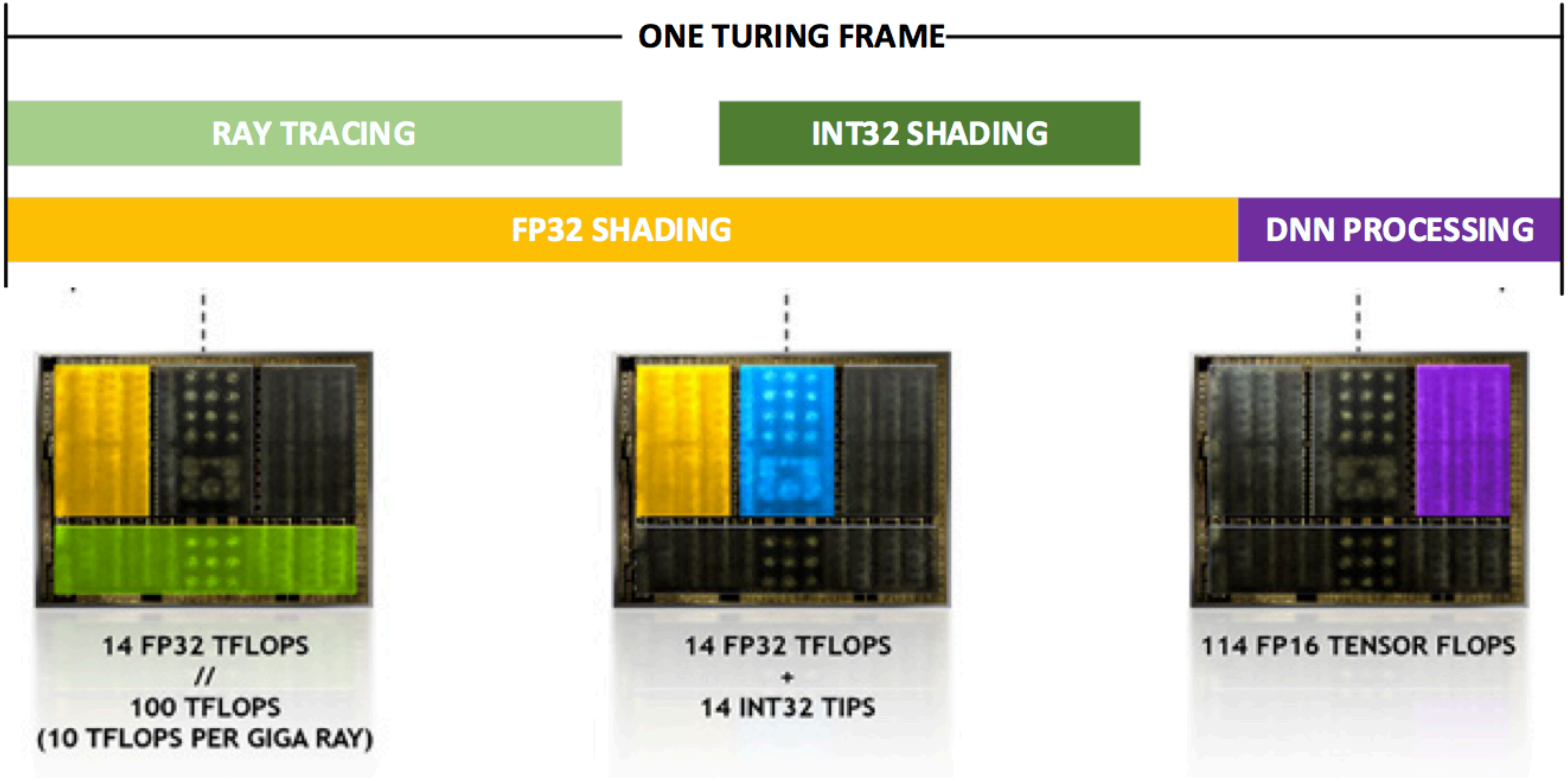
- Boldfaced numbers used for official peak performance (see next slide).
- If you do not activate GPU Boost, expect a 21% performance penalty in ALL execution times.

Performance for the flagship chip: TU102 GPU (18.6 billion transistors on TSMC’s 12nm. FFN)

Data type (accuracy)	GeForce RTX 2080 Ti Founders Edition	Quadro RTX 6000 and GeForce Titan RTX
FP32 (single precision)	68 SMs x 64 cores/SM x 1635 MHz x 1 madd = **14.2 TFLOPS**	72 SMs x 64 cores/SM x 1770 MHz x 1 madd = **16.3 TFLOPS**
FP16 (half precision)	68 SMs x 64 cores/SM x 1635 MHz x 2 madd = [21.2] **28.5 TFLOPS**	72 SMs x 64 cores/SM x 1770 MHz x 2 madd = **32.6 TFLOPS**
INT32 concurrent with FP.	14.2 TFLOPS	16.3 TFLOPS
Tensor (FP16 matrix math with FP16 accumulation)	68 SMs x 8 tensor cores/SM x 1635 MHz x 64 madds = **113.8 TFLOPS**	72 SMs x 8 tensor cores/SM x 1770 MHz x 64 madds = **130.5 TFLOPS**
Ray Tracing ops./sec.	100 Tera	100 Tera
Total throughput for a typical benchmark	78 Tera-ops	84 Tera-ops

- Total throughput = (20% Tensor cores + 80% CUDA cores) + 40% RT ops. + 28% INT32. Concurrent operations with CUDA cores: 50% Ray Tracing (40%) and 35% INT32 (28%).
- For 2080 Ti: 113.8 * 0.2 + 14.2 * 0.8 + 100 * 0.4 + 14.2 * 0.28 = 78.

Typical concurrency found on Nvidia’s benchmarks

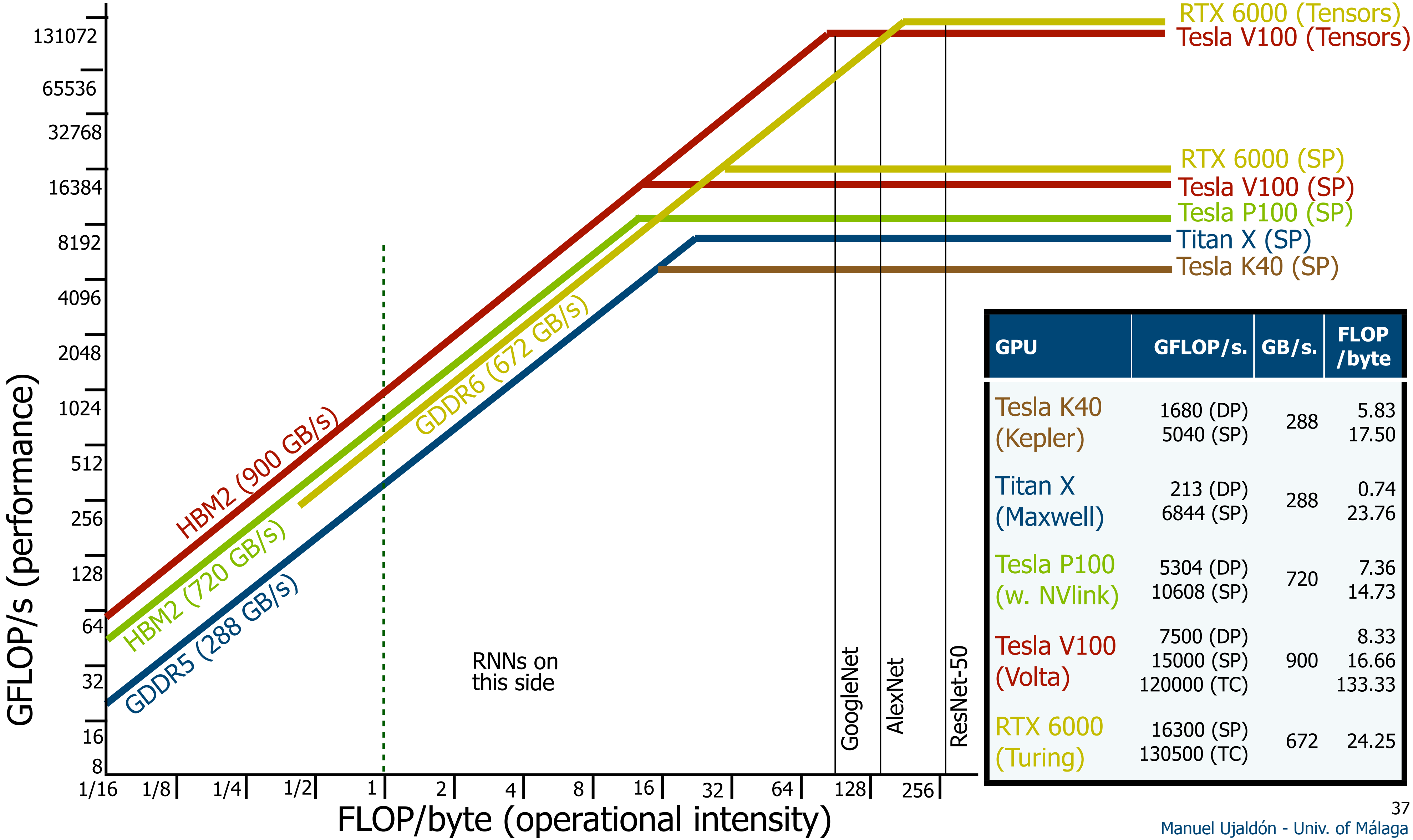


Performance for the flagship chip: TU102 GPU (18.6 billion transistors on TSMC's 12nm. FFN)

Data type (accuracy)	GeForce RTX 2080 Ti Founders Edition	Quadro RTX 6000 and GeForce Titan RTX
FP32 (single precision)	68 SMs x 64 cores/SM x 1635 MHz x 1 madd = **14.2 TFLOPS**	72 SMs x 64 cores/SM x 1770 MHz x 1 madd = **16.3 TFLOPS**
FP16 (half precision)	68 SMs x 64 cores/SM x 1635 MHz x 2 madd = [21.2] **28.5 TFLOPS**	72 SMs x 64 cores/SM x 1770 MHz x 2 madd = **32.6 TFLOPS**
INT32 concurrent with FP.	14.2 TFLOPS	16.3 TFLOPS
Tensor (FP16 matrix math with FP16 accumulation)	68 SMs x 8 tensor cores/SM x 1635 MHz x 64 madds = **113.8 TFLOPS**	72 SMs x 8 tensor cores/SM x 1770 MHz x 64 madds = **130.5 TFLOPS**
Ray Tracing ops./sec.	100 Tera	100 Tera
Total throughput for a typical benchmark	78 Tera-ops	84 Tera-ops

- Total throughput = (20% Tensor cores + 80% CUDA cores) + 40% RT ops. + 28% INT32. Concurrent operations with CUDA cores: 50% Ray Tracing (40%) and 35% INT32 (28%).
- For 2080 Ti: 113.8 * 0.2 + 14.2 * 0.8 + 100 * 0.4 + 14.2 * 0.28 = 78.

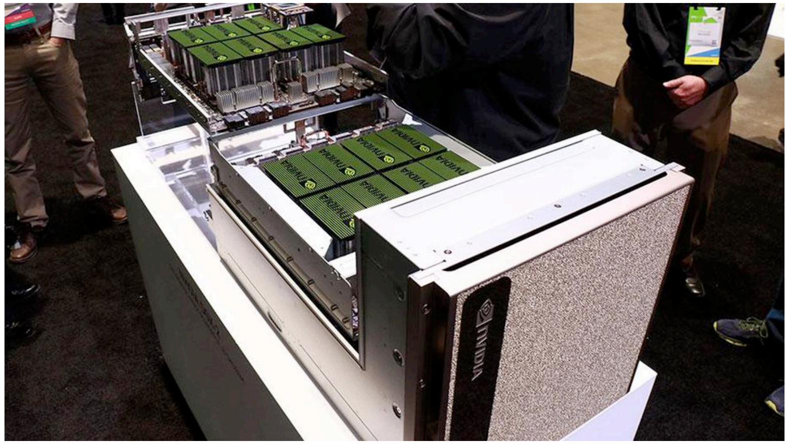
Single Precision and Deep Learning performance for the last 5 CUDA generations



GPU	GFLOP/s.	GB/s.	FLOP /byte
Tesla K40 (Kepler)	1680 (DP) 5040 (SP)	288	5.83 17.50
Titan X (Maxwell)	213 (DP) 6844 (SP)	288	0.74 23.76
Tesla P100 (w. NVlink)	5304 (DP) 10608 (SP)	720	7.36 14.73
Tesla V100 (Volta)	7500 (DP) 15000 (SP) 120000 (TC)	900	8.33 16.66 133.33
RTX 6000 (Turing)	16300 (SP) 130500 (TC)	672	24.25

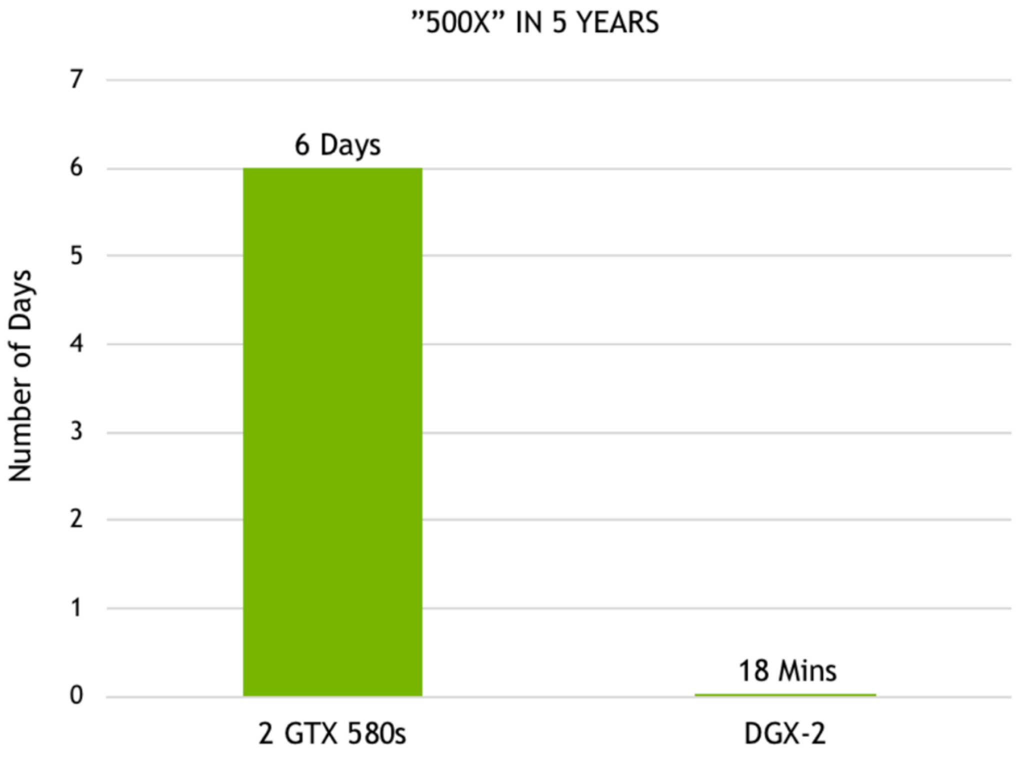
The DGX-2 supercomputer: The world largest GPU

- 16 Tesla V100 connected by NVswitch. Each switch with:
  - 2B transistors.
  - 18 links 8 bits wide.
  - 25 Gbits/sc.
  - 7.2 TB/sc (20x PCI-e 300 GB/sc.).
- On-chip memory fabric semantic extended across all GPUs.
- 512 GB. HBM2 @ 900 GB/sc [14.4 TB/sc. aggregate].
- 10.000 watts.
- 158 kilograms.
- 10x DGX-1 (sept’17).
- 399.000 USD.

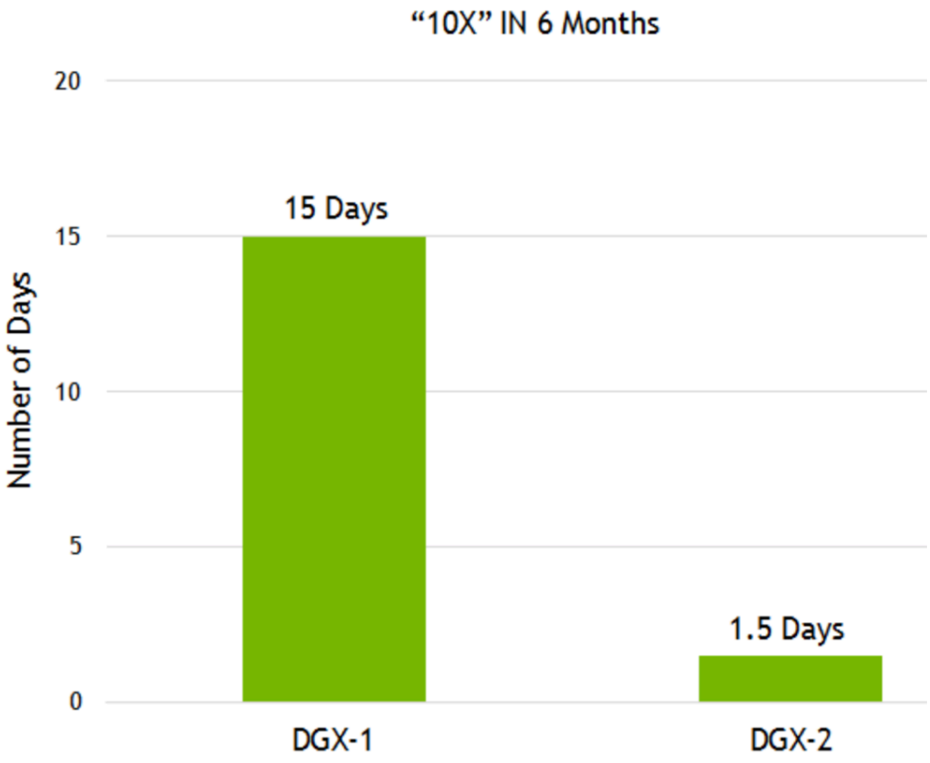


Performance evolution using the DGX-2 supercomputer

Time to train AlexNet on the Imagnet Dataset



Time to train Facebook’s Fairseq



Latest HPC performance achievements

- For the first time in history, most of the FLOPS added to the top500.org list came from GPUs.
- Taken together, the 3 GPU supercomputers in the top5 represent more deep learning capability than the other 497 systems ranked in the top500 list.
- 97% of the Summit peak performance is derived from its 26.136 GPUs.
- Tensor cores deliver 120-130 TFLOPS (peak) for DL.
- Most applications do not exploit more than 5% of peak performance in modern HPC supercomputers [1.3-1.8 HPCG].
- There are already more than 500 popular scientific codes ported to Volta/Turing, including all of the top 15 HPC apps.

Concluding remarks

- The Volta GPU accelerates graphics, HPC and AI, enabling data scientist, researchers and engineers to tackle grand-challenge applications in an unprecedented way.
- Welcome to the dark silicon era, introducing **meta-chips**: A 20+ billion transistors chip can afford to contain a bunch of sub-chips, each aiming to a different user profile.
- Deep Learning users represent applications that hardware architects always wanted to have: compute bound!
- GPU processing becomes the main trend in HPC for the first time in HPC history.

Thanks so much for your attention

- You can always reach me in Spain at the Computer Architecture Department of the University of Malaga:
  - e-mail: ujaldon@uma.es
  - Phone: +34 952 13 28 24.
  - Web page: http://manuel.ujaldon.es (english/spanish versions available).

- QUESTIONS?