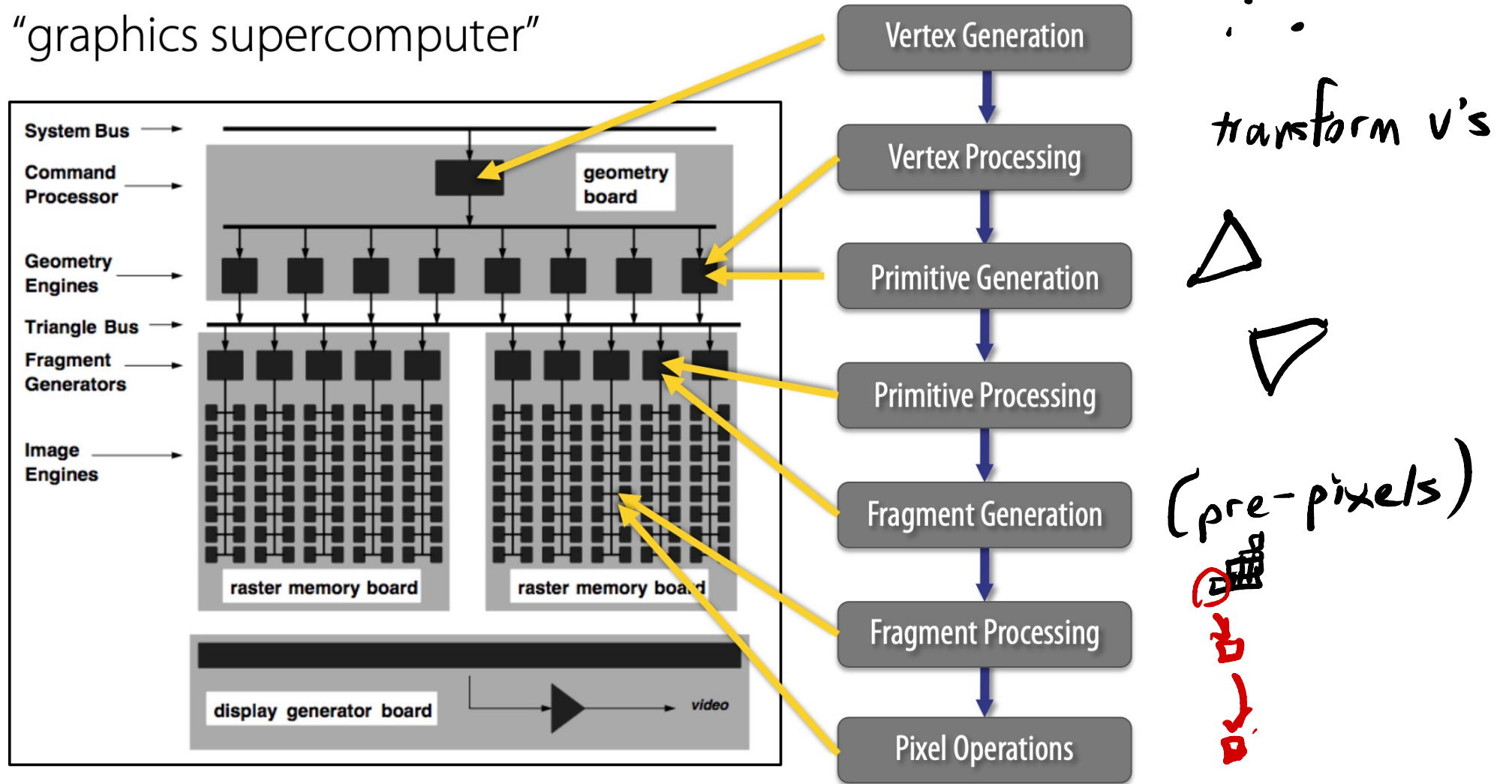


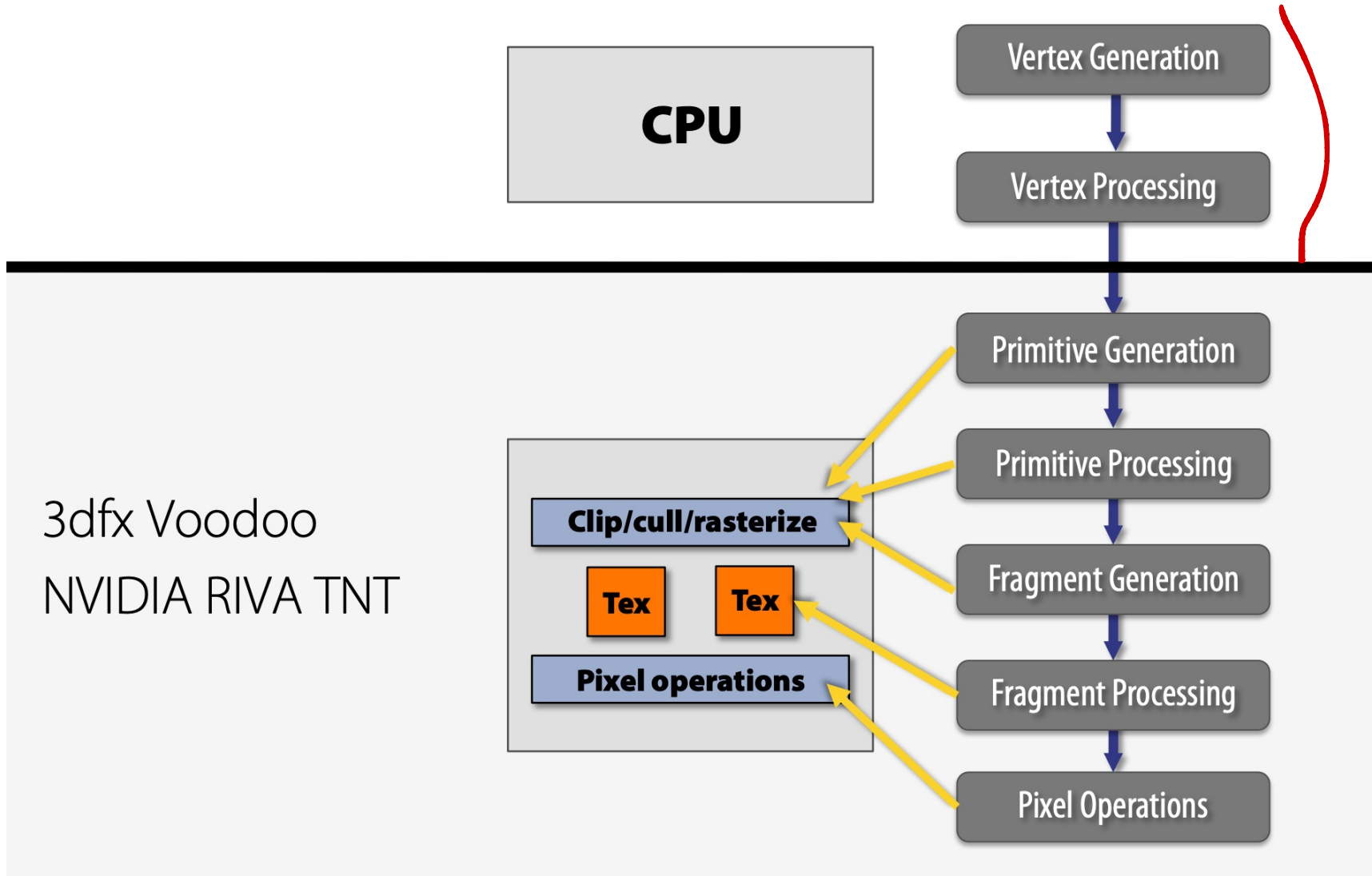
17 – gpu

Silicon Graphics RealityEngine (1993)

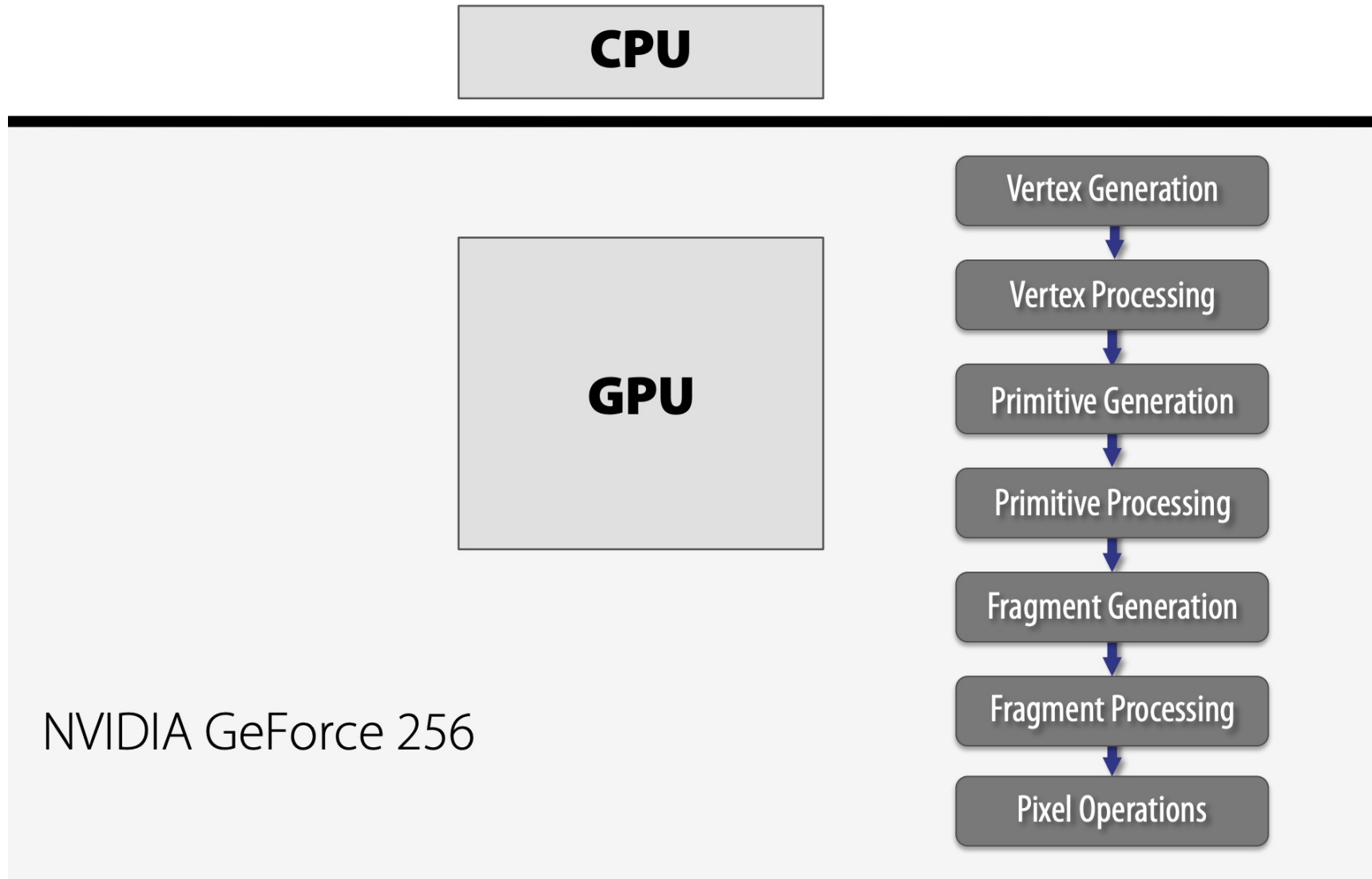
“graphics supercomputer”



Pre-1999 PC 3D graphics accelerator



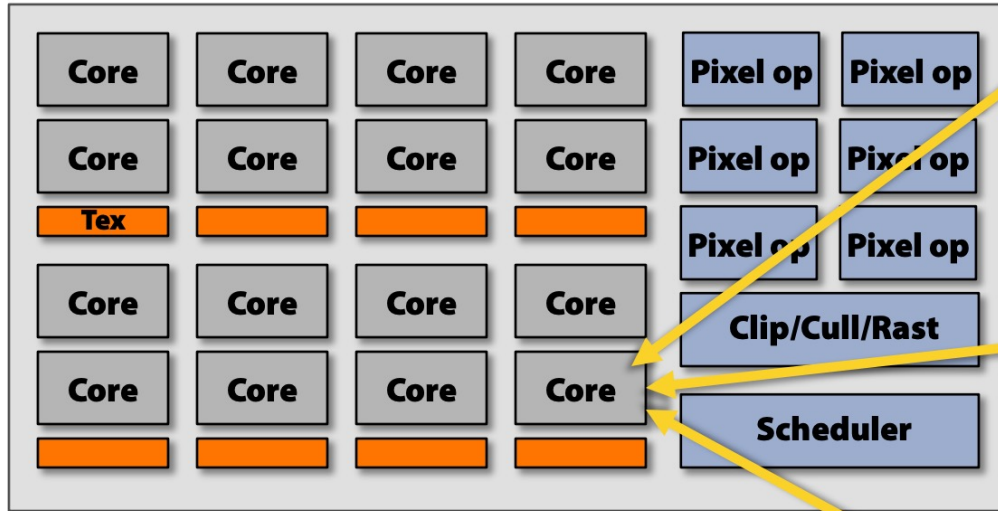
GPU* circa 1999



Direct3D 10 programmability: 2006

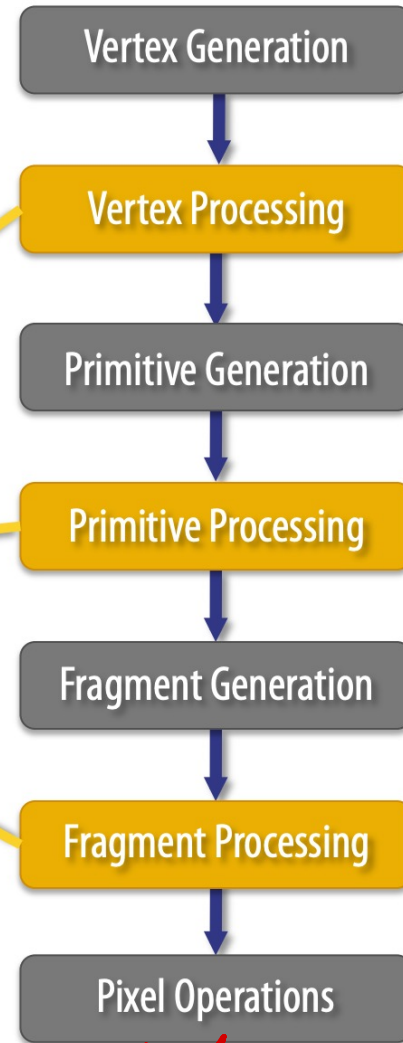
unified architecture

General Processors



NVIDIA GeForce 8800
("unified shading" GPU)

start of "GPGPU" experimentation



up to here we used "fixed" pipeline

eg. matrix stack

& start of shader programming

Nvidia GPUs

	2006 Tesla	2009 Fermi	2012 Kepler	2014 Pascal	2018 Turing	2020 Ampere
cores	128	512	2880	3584	4352	6912
transistors	681 million	3 billion	7.1 billion	12 billion	18.6 billion	54.2 billion
clock	1.5 Ghz	2 Ghz	0.89 GHz	1.4 / 1.6 GHz	1.3 / 1.8 GHz	1.5 / 1.7 GHz
perf (32)	576 Gflops	1.5 Tflops	5 Tflops	11.3 Tflops	14.2 Tflops	19.5 Tflops
fp16					30 Tflops	78 Tflops
tensor16					125 Tflops	312 Tflops
tensor32/64						156 / 19.5 Tflops
examples	Geforce 8 - 300	GeForce 400-500	GeForce 600-700	GTX 1000 series	GTX 1600 / RTX 2000	RTX 3000

Maxwell (2014) is GTX 750-980 series

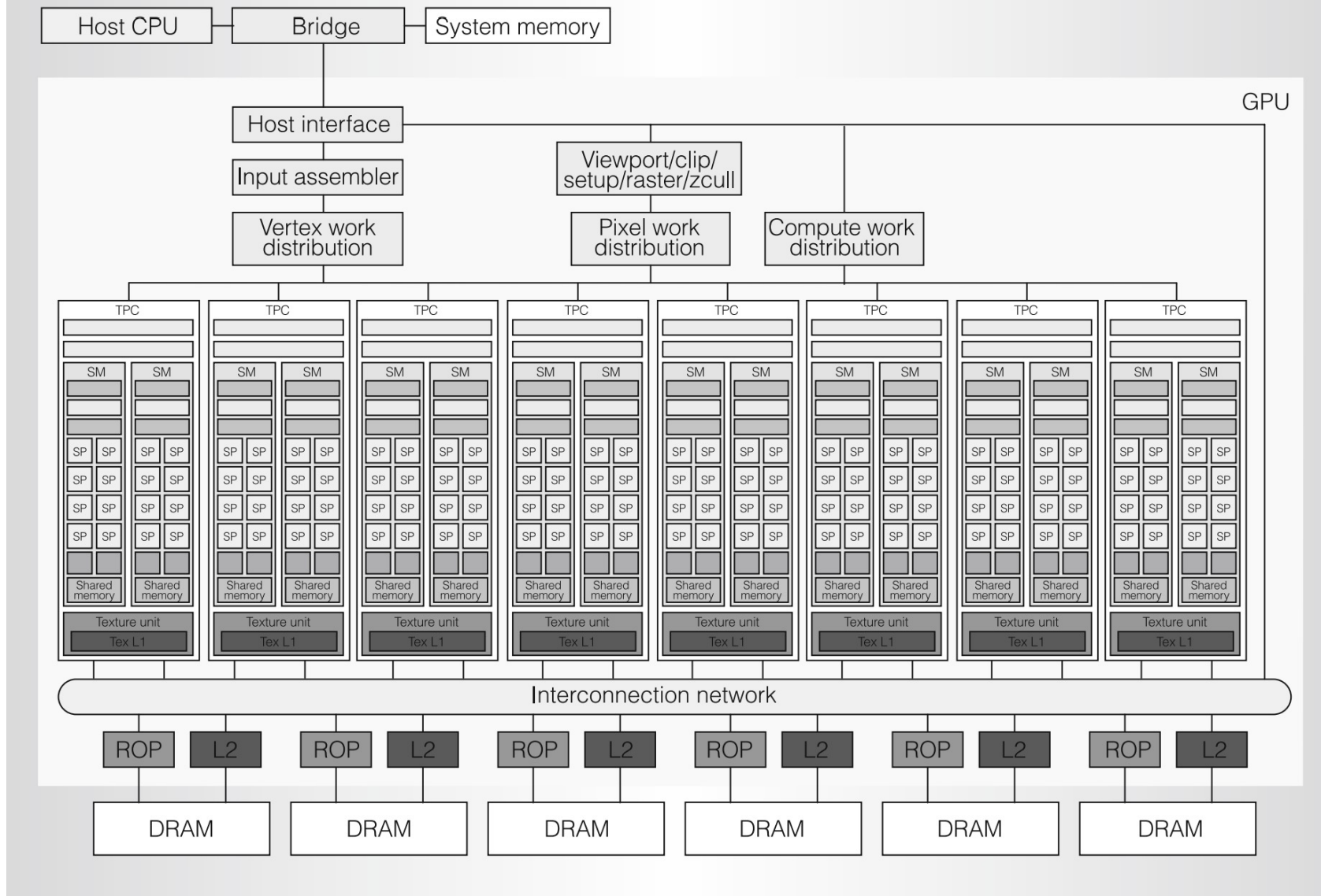
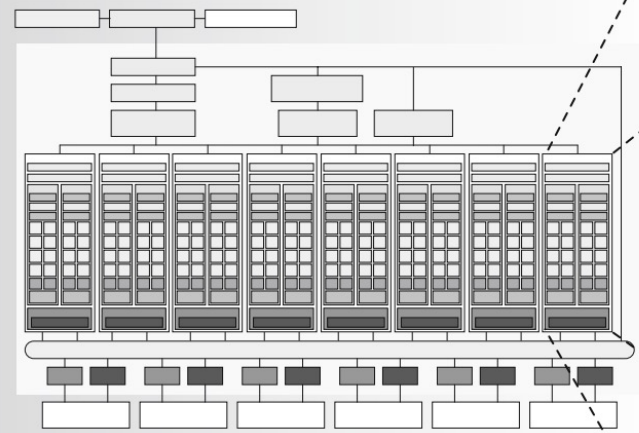
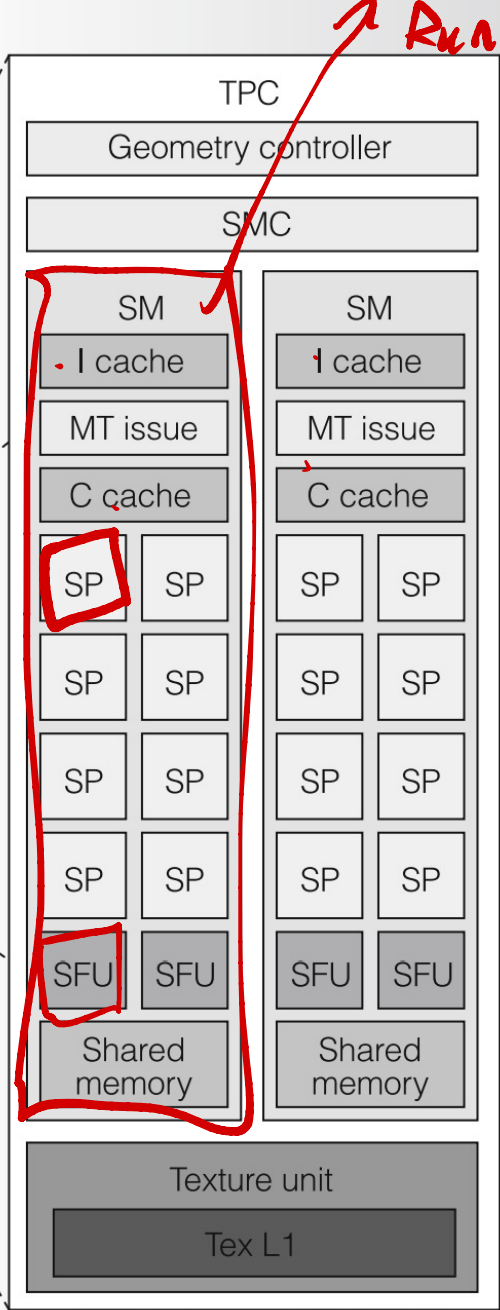


Figure 1. Tesla unified graphics and computing GPU architecture. TPC: texture/processor cluster; SM: streaming multiprocessor; SP: streaming processor; Tex: texture, ROP: raster operation processor.

SP - streaming processor



SFU Special function unit
trig, $\frac{1}{x}$, $\frac{1}{\sqrt{x}}$
log.



Run in parallel same pgm

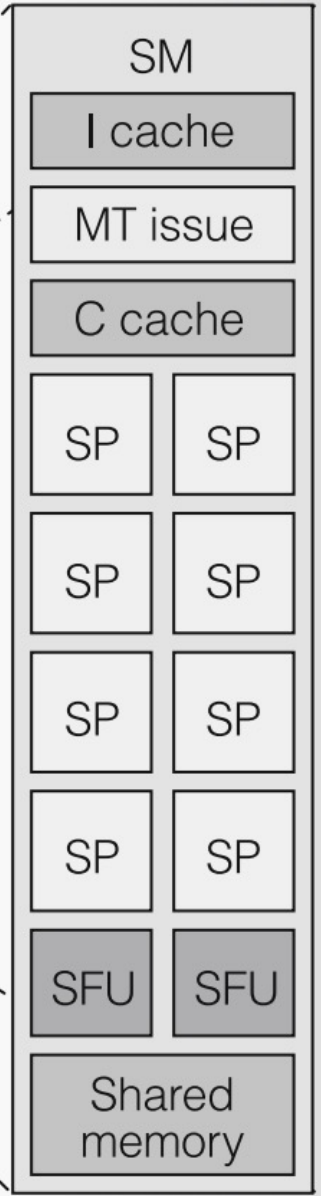
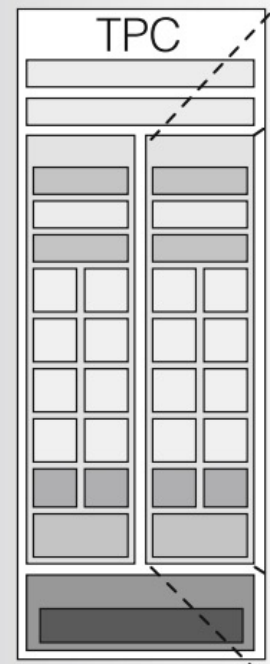


Figure 2. Texture/processor cluster (TPC).

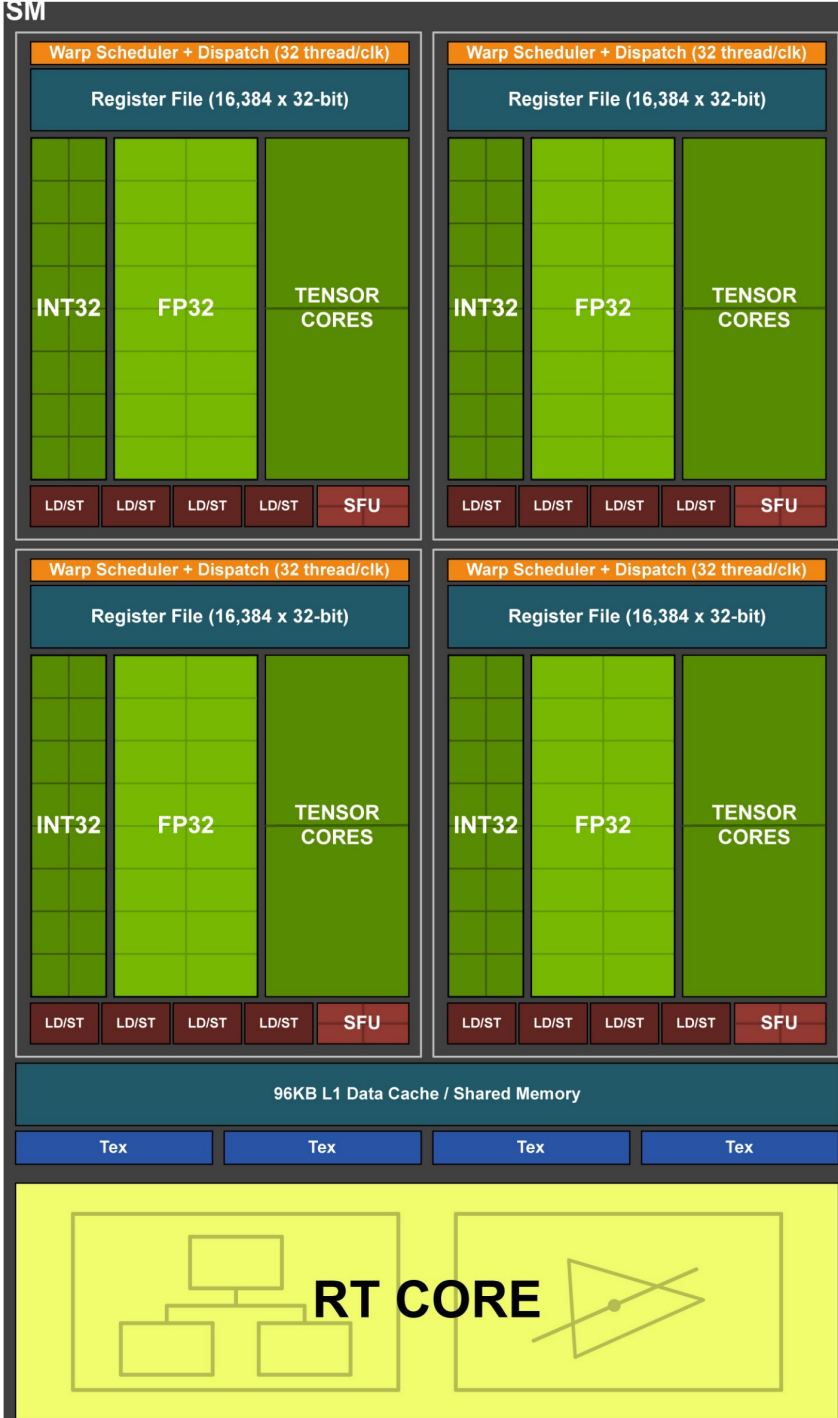
Figure 3. Streaming multiprocessor (SM).

NC-SA 4.0)

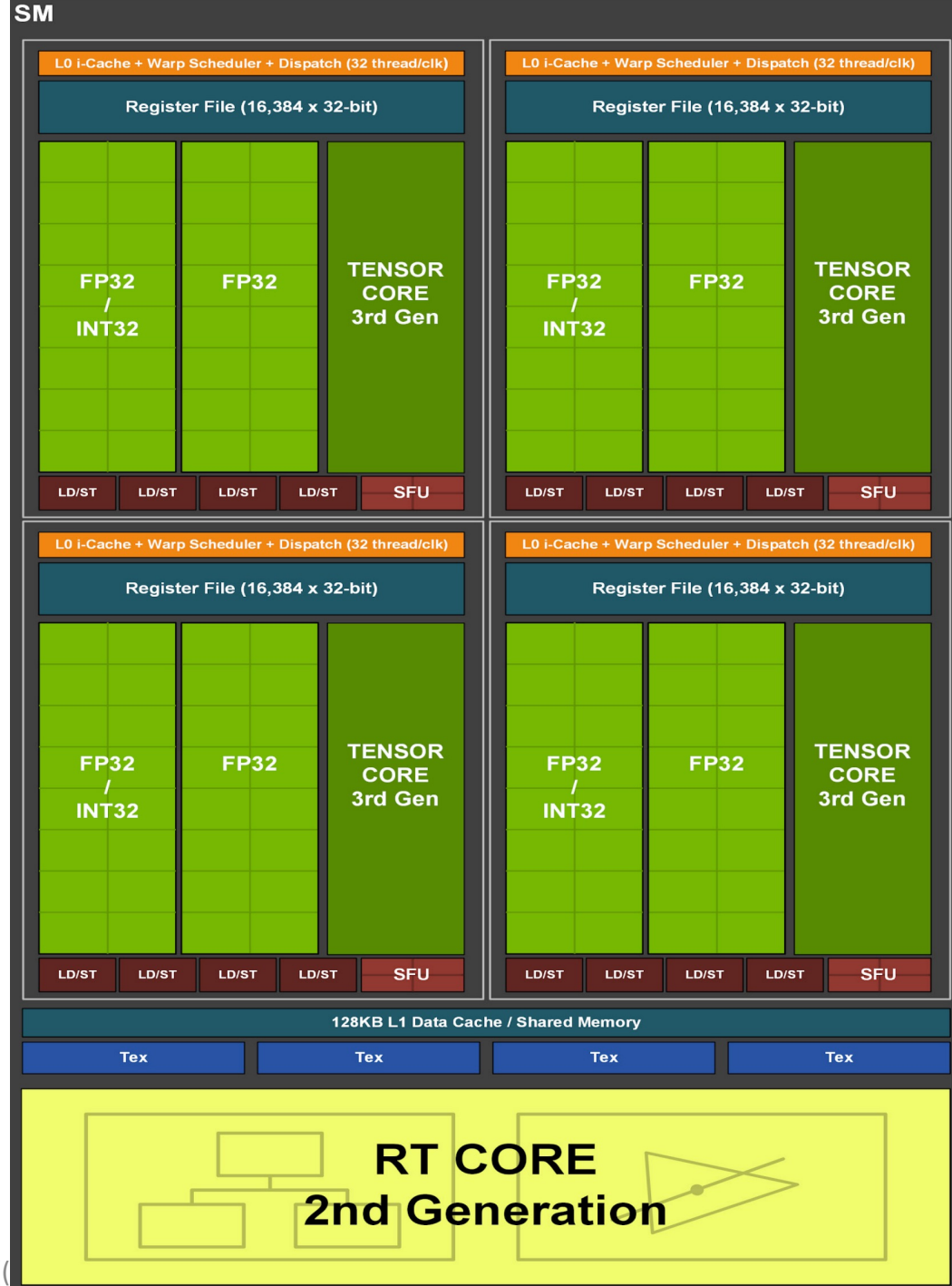


Turing architecture

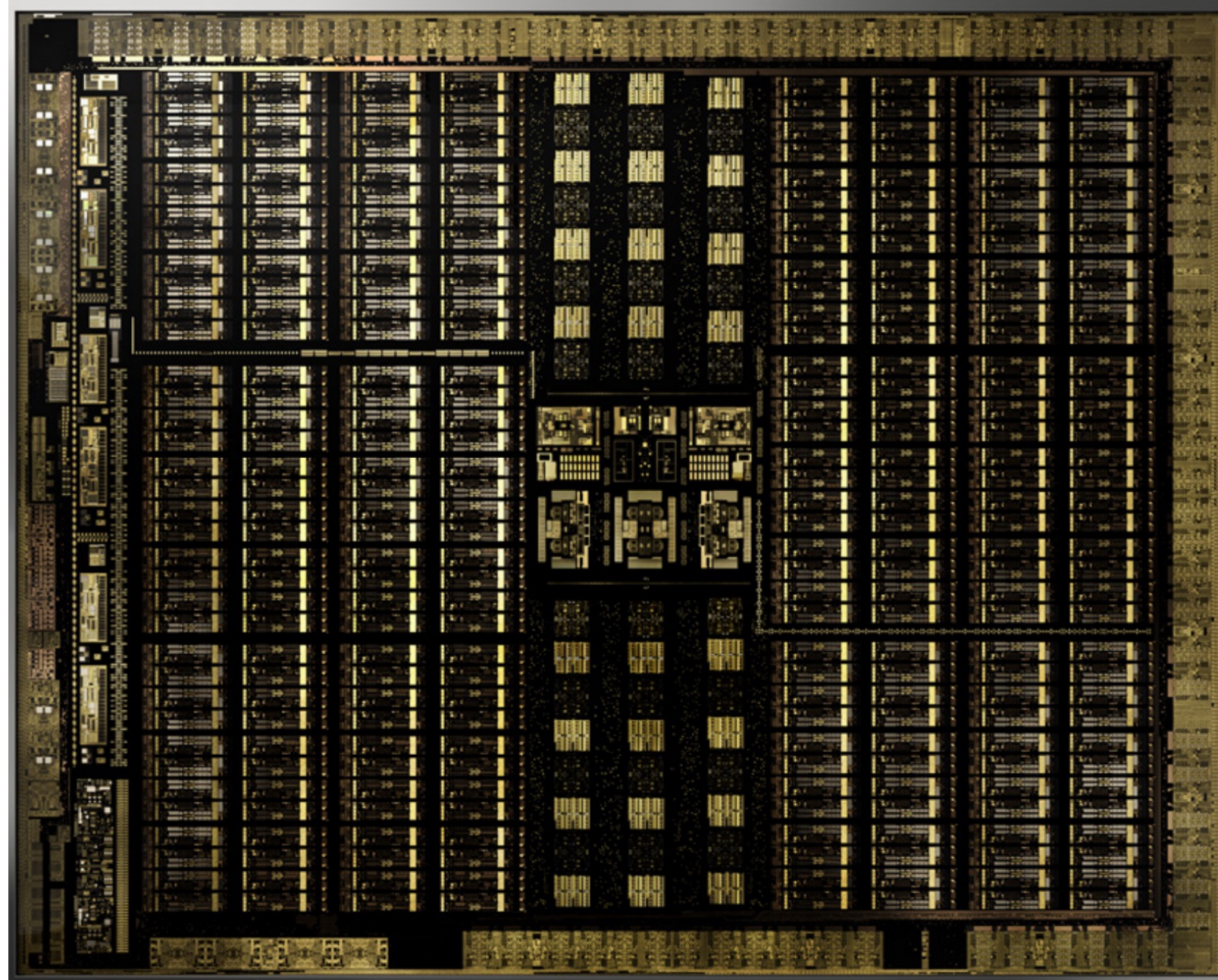




Turing
vs
Ampere



Turing Chip



Turing Arch

4608 CUDA cores

72 Ray tracing core

576 tensor cores (neural net / ML / ...)

bounding volume hierchies (box)

ray / triangle intersection

