Turing Architecture

Gahan M. Saraiya (18MCEC10)

M.Tech (Computer Science and Engineering)
Institute of Technology, Nirma University, Ahmedabad

November 2018

Outline of Talk

What is Turing Architecture? I

- codename for a GPU microarchitecture developed by Nvidia
- Successor of Volta
- named after well known mathematician and computer scientist Alan Turing
- the first consumer product with capability of real-time ray-tracing
- NVIDIA partnered with Microsoft to enable full RTX support via Microsoft's new DirectX Raytracing (DXR) API
- TU102 GPU includes **18.6 billion transistors** fabricated on TSMC's 12 nm FFN (FinFET NVIDIA) high-performance manufacturing process

What is Turing Architecture? II

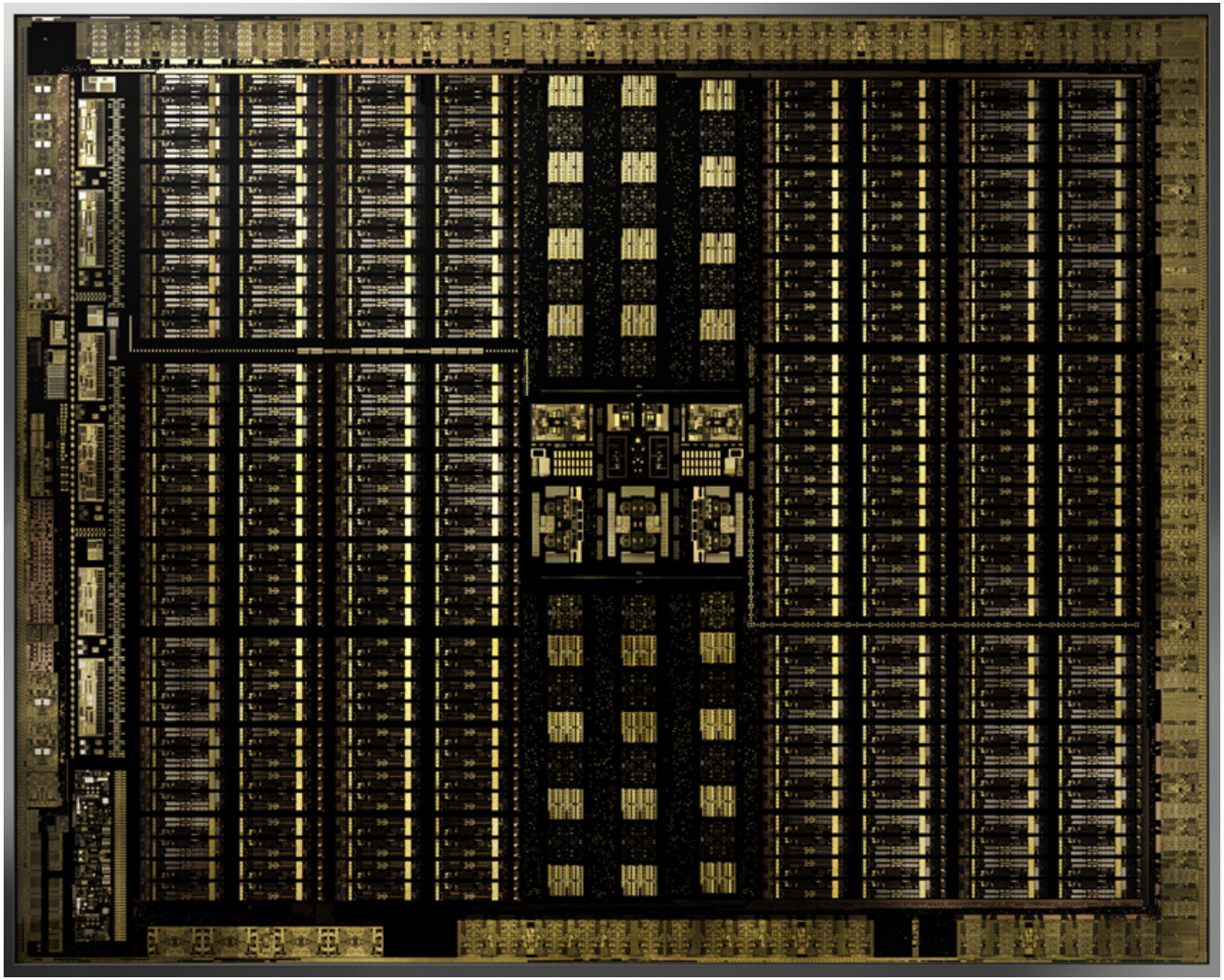


Figure: Turing Architecture in TU102

Overview of Product

Figure: RTX 2080 Ti

14.2 TFLOPS[1] of peak single precision (FP32)
113.8 Tensor TFLOPS
10 Giga Rays/sec
78 Tera RTX-OPS

Figure: QUADRO RTX 6000

16.3 TFLOPS[1] of peak single precision (FP32)
130.5 Tensor TFLOPS
10 Giga Rays/sec
84 Tera RTX-OPS

[1] Based on GPU boost clock

Outline of Talk

The New Features I

RT Cores for Real-Time Ray Tracing

helps single GPU to render realistic **cinematic-quality 3D games, complex professional models** with accurate shadows, reflections, refractions.
The Turing architecture is armed with dedicated ray-tracing processors called RT Cores that accelerate the computation of how light and sound travel in 3D environments by up to 10 Giga Rays per second.

Tensor Cores for AI Acceleration

Tensor Cores - processors that accelerate deep learning training and inference, providing up to **500 trillion tensor operations per second**.
This level of performance dramatically accelerates AI-enhanced features—such as denoising, resolution scaling, and video re-timing—creating applications with powerful new capabilities.

The New Features II

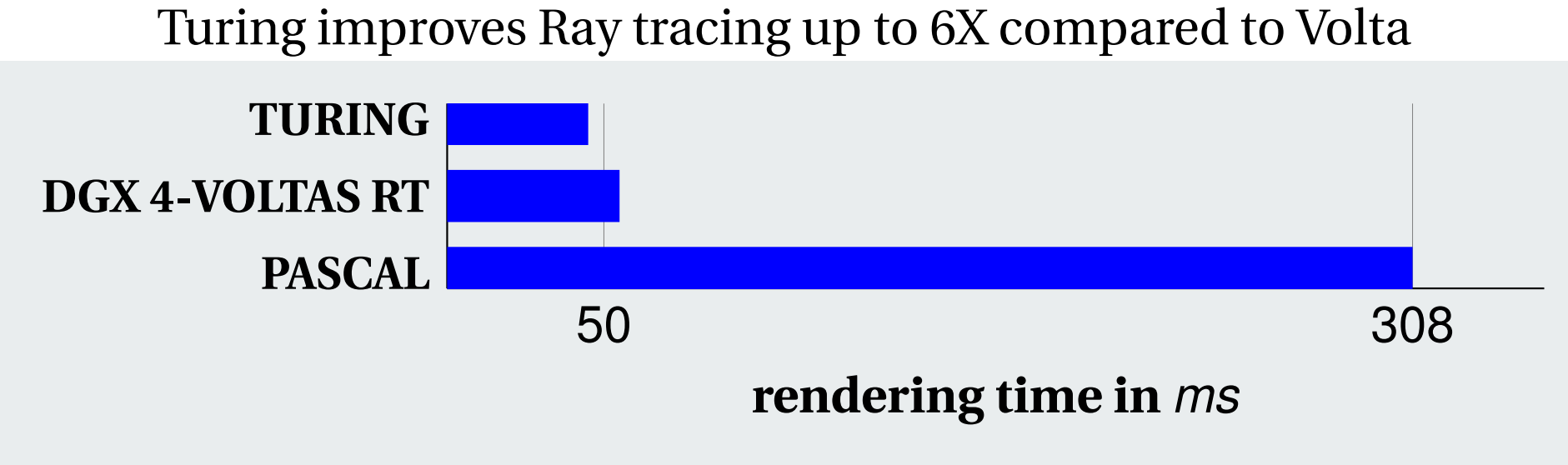
New Streaming Multiprocessor

The Turing architecture dramatically improves raster performance over the previous-generation Pascal with an enhanced graphics pipeline and new programmable shading technologies (variable-rate shading, texture-space shading, and multi-view rendering).

CUDA For Simulation

Turing-based GPUs feature a new streaming multiprocessor (SM) architecture that supports **up to 16 trillion floating-point operations** in parallel with 16 trillion integer operations per second.
Developers can take advantage of up to 4,608 CUDA cores with NVIDIA CUDA 10, FleX, and PhysX software development kits (SDKs) to create complex simulations, such as particle or fluid dynamics for scientific visualization, virtual environments, and special effects.

Ray Tracing


Turing improves Ray tracing up to 6X compared to Volta
TURING
DGX 4-VOLTAS RT
PASCAL
50
308
rendering time in ms

Comparing Workstation GPU I

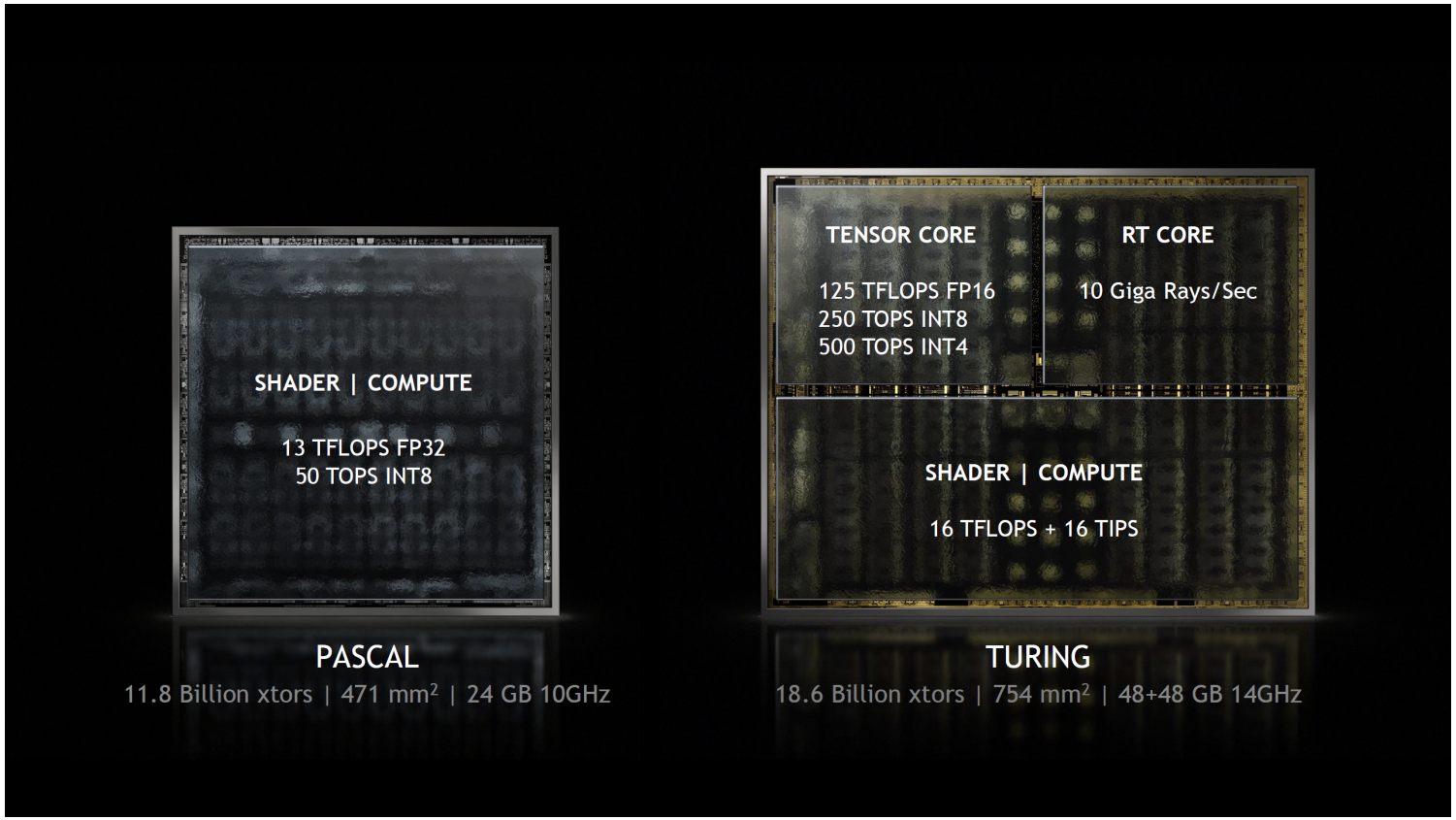
	Quadro RTX 6000	**Volta 100**
CUDA Parallel-Processing Cores	4608	5120
Architecture	Turing	Volta
Code Name	TU102	GV100
Transistor count	18,600 million	21,100 million
Transistors	12nm	12nm
Core clock speed	1440 MHz	1132 MHz
Memory clock speed	12000 MHz	1696 MHz

Comparing Workstation GPU II

Tensor Cores	576	640
RT Cores	72	-
GPU Memory	24 GB GDDR6	32 GB HBM2
RTX-OPS	84T	-
Rays Cast	10 Giga Rays/Sec	-
FP32 Performance	16.3 TFLOPS	14.8 TFLOPS
Tensor Performance	16.3 TFLOPS	130.5 TFLOPS
Max Power Consumption	295 W	250 W

Comparing Workstation GPU III

Graphics Bus	PCI Express 3.0 x 16	PCI Express 3.0 x 16
Display Connectors	DP 1.4 (4), VirtualLink (1)	DP 1.4 (4)
Form Factor	4.4" H × 10.5" L Dual Slot	4.4" H × 10.5" L Dual Slot
Release date	13 August 2018	27 March 2018
Launch Price	$6299	$8999


SHADER | COMPUTE
13 TFLOPS FP32
50 TOPS INT8
PASCAL
11.8 Billion xtors | 471 mm² | 24 GB 10GHz
TENSOR CORE
125 TFLOPS FP16
250 TOPS INT8
500 TOPS INT4
RT CORE
10 Giga Rays/Sec
SHADER | COMPUTE
16 TFLOPS + 16 TIPS
TURING
18.6 Billion xtors | 754 mm² | 48+48 GB 14GHz

Outline of Talk

Tensor Core I

- Tesla T4 introduces NVIDIA Turing Tensor Core technology with multi-precision computing for the world's most efficient AI inference.
- Turing Tensor Cores provide a full range of precisions for inference, from FP32 to FP16 to INT8, as well as INT4, to provide giant leaps in performance over NVIDIA Pascal® GPUs.
- however it's predecessor Nvidia V100 GPU (Volta) having first-generation Tensor Cores can only deliver performance with mixed-precision matrix multiply in FP16 (12X compare to PASCAL) and FP32(6X compare to PASCAL)

Tensor Core II

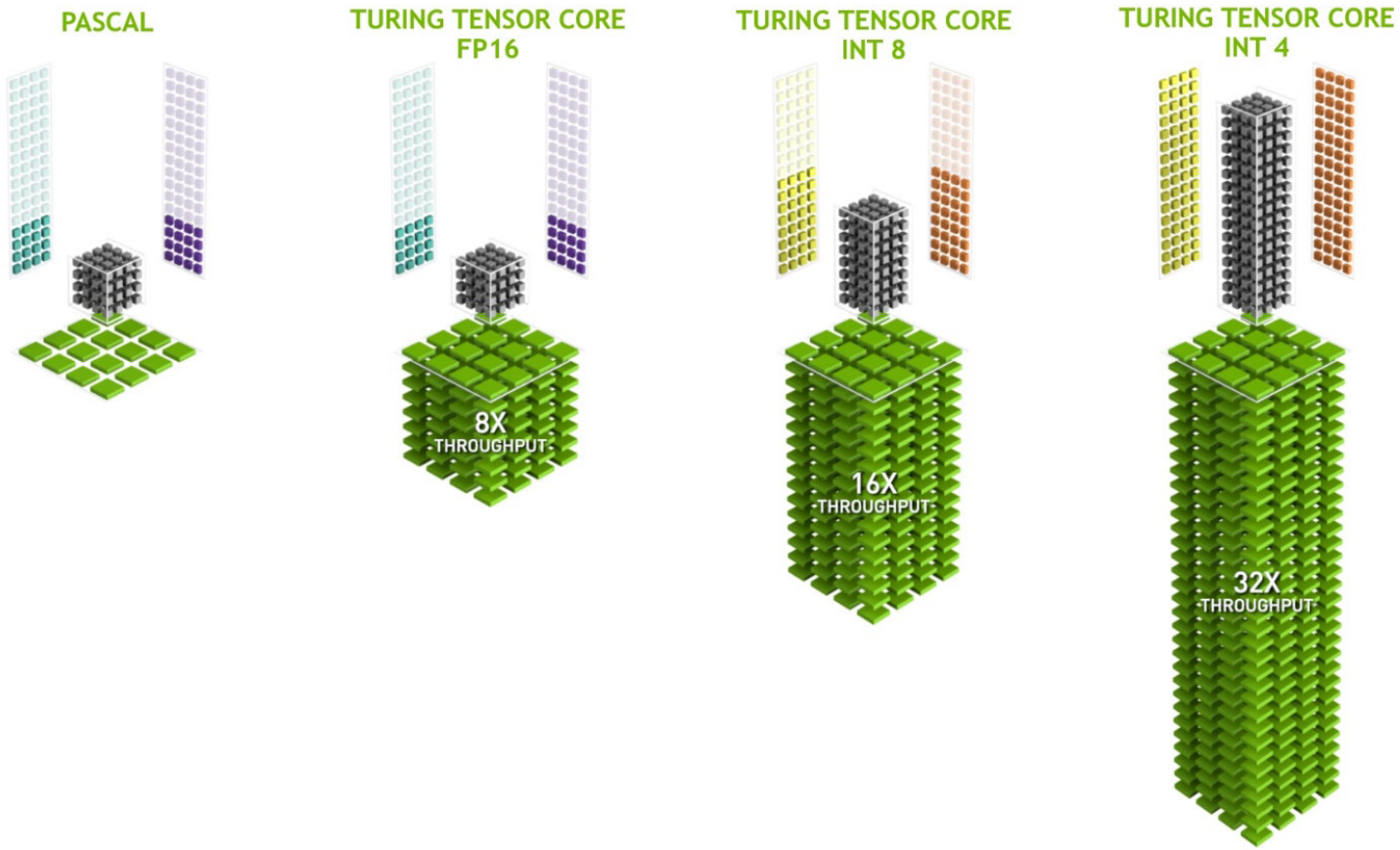


Figure: Turing Tensor core throughput comparison

Outline of Talk

Terminologies I

- Rasterization

  From a mesh of virtual triangles or polygons objects on the screen are created which indeed create 3D models of objects. In this virtual mesh, the corners of each triangle — known as vertices — intersect with the vertices of other triangles of different sizes and shapes.

- Real-Time Ray Tracing

  helps single GPU to render realistic cinematic-quality 3D games, complex professional models with accurate shadows, reflections, refractions.

- RT Core

  helps single GPU to render realistic 3D games, complex professional models with accurate shadows, reflections, refractions.

Terminologies II

- **Tensor Core**
  can accelerate large matrix operations
  perform mixed-precision matrix multiply and accumulate calculations in a single operation

  With hundreds of Tensor Cores operating in parallel in one NVIDIA GPU, this enables

  massive increases in throughput and efficiency