



NVIDIA.

## Ray Tracing vs. Rasterization

Is this the right question to ask?

## History of Predicting the Future

- I think there is a world market for maybe 5 computers. - Thomas J. Watson, 1943.
- I can assure you on the highest authority that data processing is a fad and won't last out the year. - A Prentice-Hall editor, 1957.
- There is no reason anyone would want a computer in their home. - Ken Olson, DEC, 1977.
- I could never render images using hardware - anonymous Animator, 2001.



## What do they use? Pencils?

- CPUs are hardware, too
  - Hardware is necessary to run SW - programmable HW
  - SW renderers runs on CPUs
- GPUs are programmable HW
  - Can run SW on GPUs now
  - Cg is a high-level language for GPUs
- GPUs are already "Turing complete" processors
  - !!! can compute any image that a CPU can !!!
- All that we need is
  - Faster GPUs – double perf every 6 months
  - More flexible GPUs – more programmability



## nVIDIA Historicals

Season	Product	Fill rate	Yr rate	Tri rate	Yr rate
2H97	Riva 128	20M	-	3M	-
1H98	Riva ZX	31M	2.4	3M	1.0
2H98	Riva TNT	50M	2.6	6M	4.0
1H99	TNT2	75M	2.3	9M	2.3
2H99	GeForce	120M	2.6	15M	2.8
1H00	GeForce2	200M	2.6	25M	2.8
2H00	GF2 Ultra	250M	1.6	31M	1.5
1H01	GeForce3	500M	4.0	30M	1.0
2H01	GeForce4	1000M	4.0	75M	5.0
			2.75		2.55

Yearly Growth well above Moore's Law



## Semiconductor Scaling Rates

Parameter	Current Value	Yearly Factor	Years to Double (Half)
Moore's Law (grids on a die)**	1 B	1.49	1.75
Gate Delay	150 pS	0.87	(5)
Capability (grids / gate delay)		1.71	1.3
Device-length wire delay		1.00	
Die-length wire delay / gate delay		1.71	1.3
Pins per package	750	1.11	7
Aggregate off-chip bandwidth		1.28	3

\*\* Ignores multi-layer metal, 8-layers in 2001

From: *Digital Systems Engineering*, Dally and Poulton



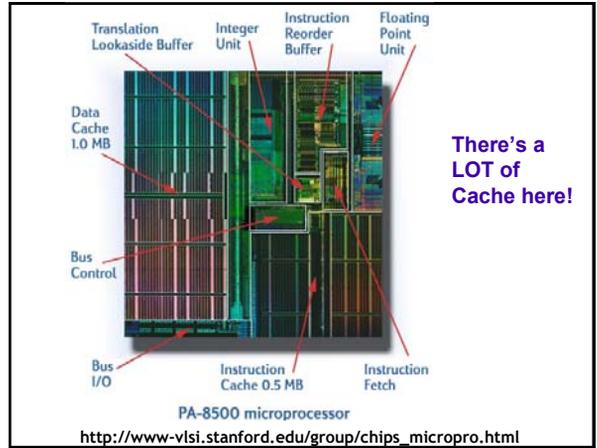
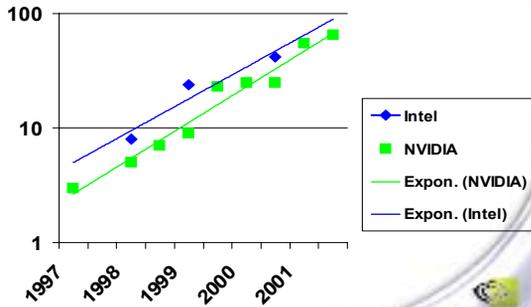
## GPU vs. CPU

- Semiconductor "Capability" grows at ~1.7x per year
- CPU performance compounding at ~1.4-1.5x per year (not bad, but wasting .2-.3x EVERY year)
- GPU performance compounding at 2.5 – 2.75x per year

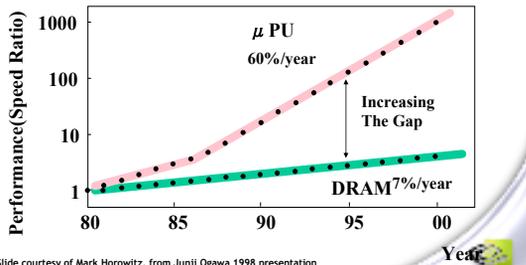
## Why is this?



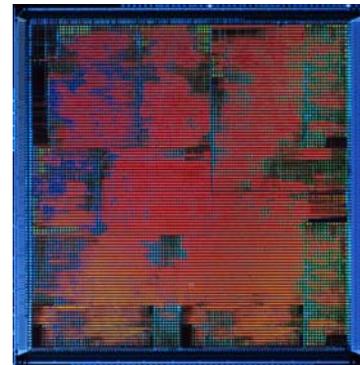
## Millions of Transistors vs. Year



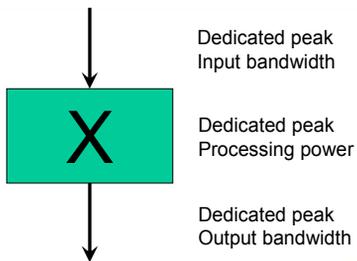
## Speed Gap between DRAM and CPU - Memory Wall



## NVIDIA GeForce3 – almost ALL logic



## “Stream” Processors



## Now: The GeForce3/GeForce4/XBOX Graphics Pipeline

vertex	programmable vertex processing
polygon	polygon setup & rasterization
pixel	per-pixel interpolation
texture	programmable per-pixel math, blending
image	Z-buffer, blending & anti-alias

## The Next Generation – Cinematic Realism in Real-time

- Full 32-bit Precision IEEE Floating Point, throughout the pipeline
- Much more programmability
  - Longer shader programs
  - Branching/looping
- Much more parallelism and shading IPS
- Still some hardwired components:
  - Rasterization, Texture Filtering, Z-buffer/Stencil/Blend
  - Hardwired logic is FAST and SMALL



## Algorithms Running on GPUs

- Ray Tracing (you've seen it!)
- Image Based Rendering
- Volume Rendering
- Radiosity (form factors, gathering)
- Using
  - Programmable elements
  - Hardwired elements, too
- So, is it really Ray Tracing vs. Rasterization?

