CS 152
Computer Architecture and Engineering
Lecture 23 -- GPU + SIMD + Vectors II

2014-4-17

John Lazzaro

(not a prof - "John" is always OK)

TA: Eric Love

www-inst.eecs.berkeley.edu/~cs152/

Play:

Today: Architecture and Graphics

- **Time Machine:** Building expensive prototypes to see the future of research.
- **ClipMaps:** How does Earth view in Google Maps cache 25 PetaBytes?
- **SGI RealityEngine:** 1993's $1M+ time machine for the Maps prototype

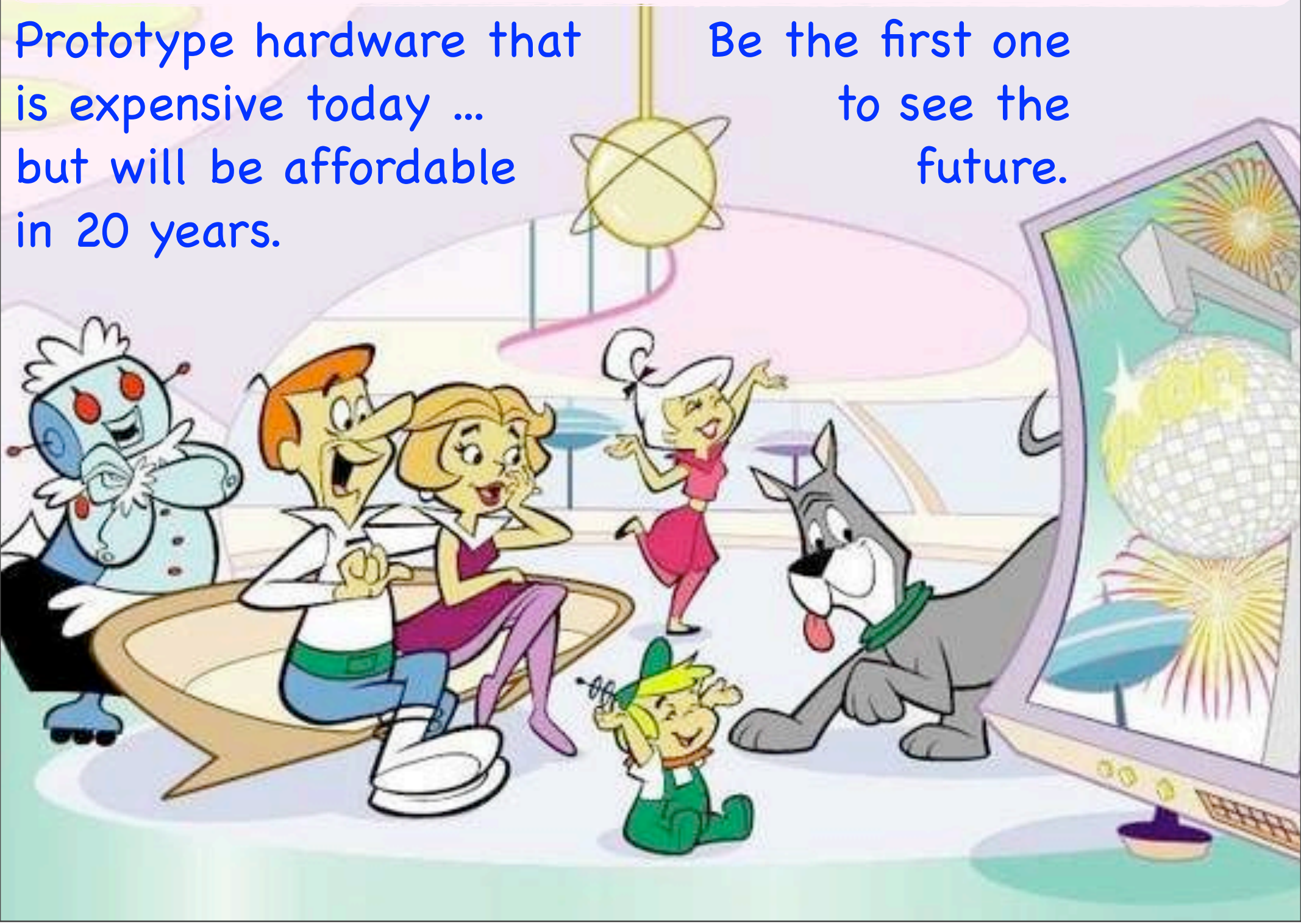
Short Break

- **Future to Present:** The path from the SGI Reality Engine to Nvidia GPUs.

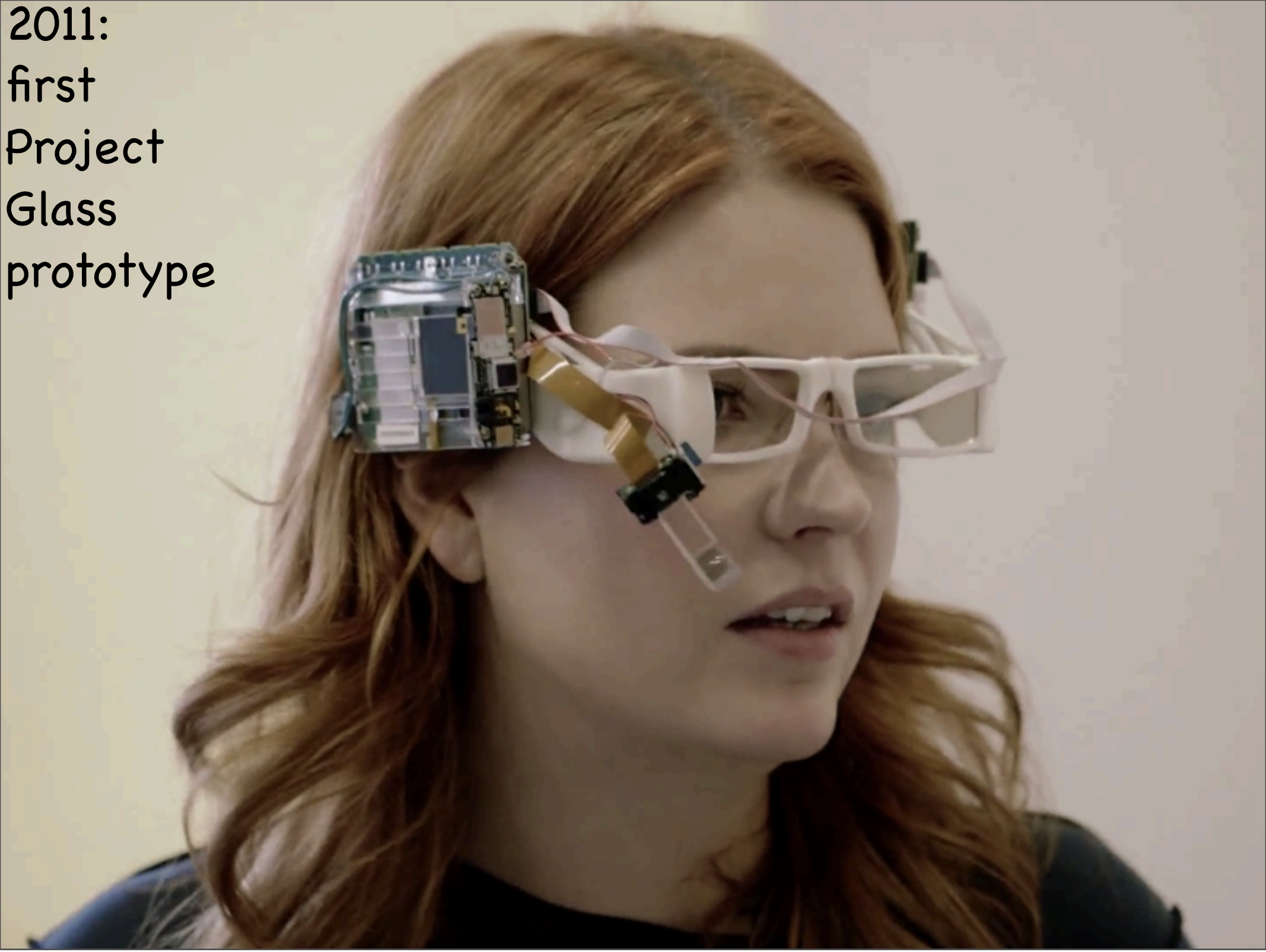
The time machine model of computer science research.

Prototype hardware that is expensive today ... but will be affordable in 20 years.

Be the first one to see the future.



2011:
first
Project
Glass
prototype

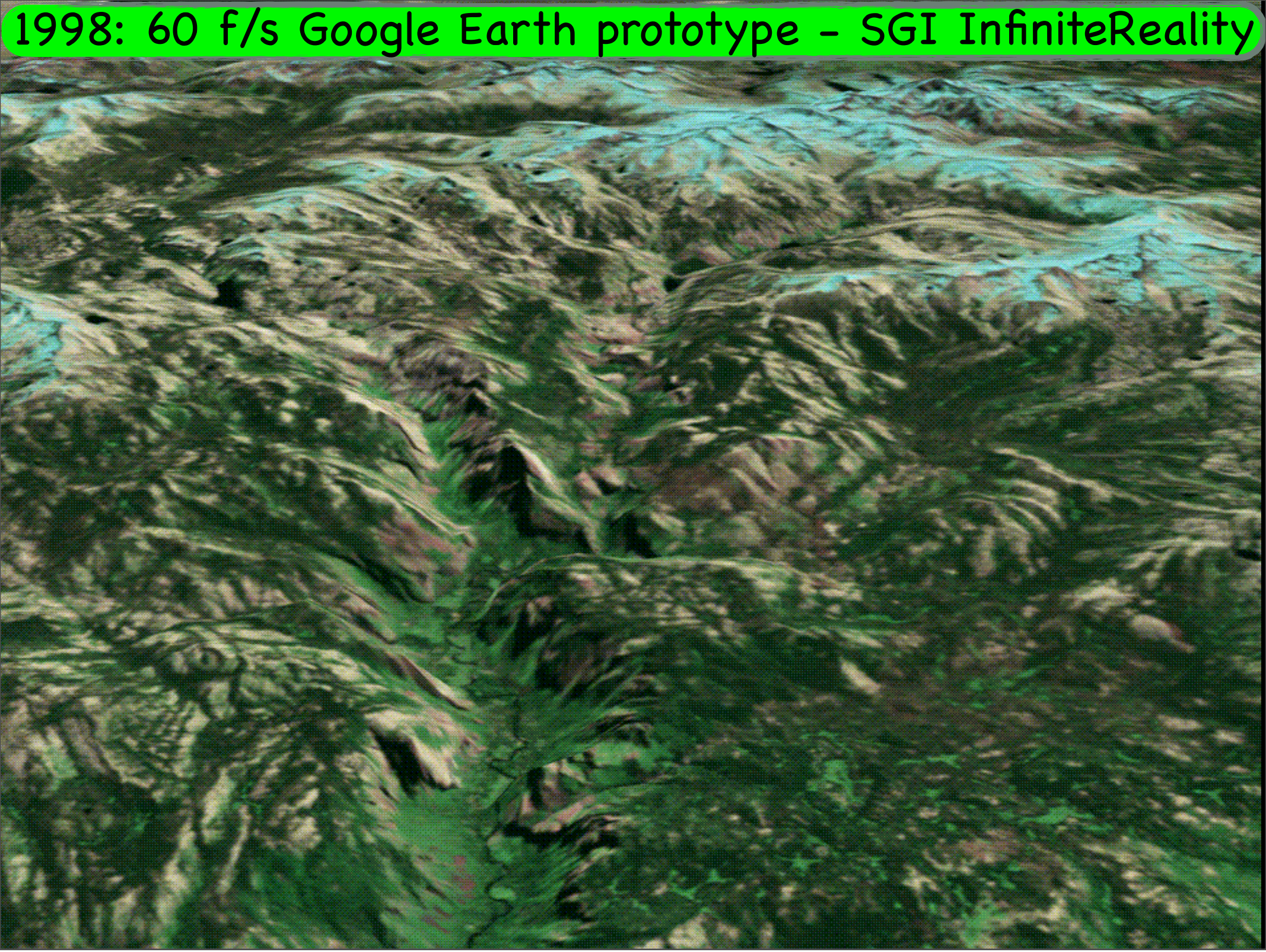


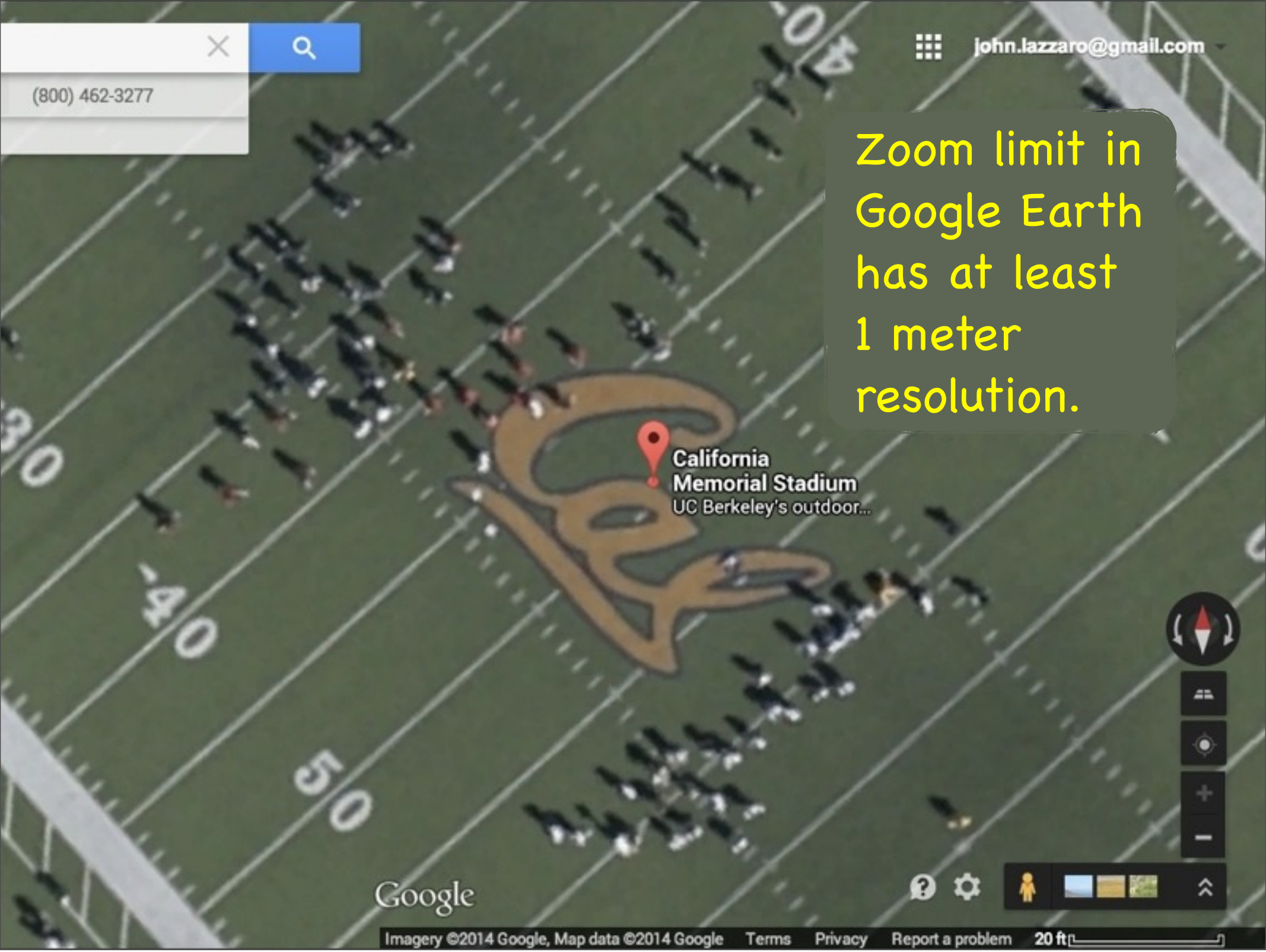
2012:
Glass
Explorer
prototype

1993 example: Walk through of a city, at 30 frames/sec.
A $1.5M visualization supercomputer (SGI RealityEngine)

Like Google Street View, but hand-crafted city models.

1998: 60 f/s Google Earth prototype - SGI InfiniteReality



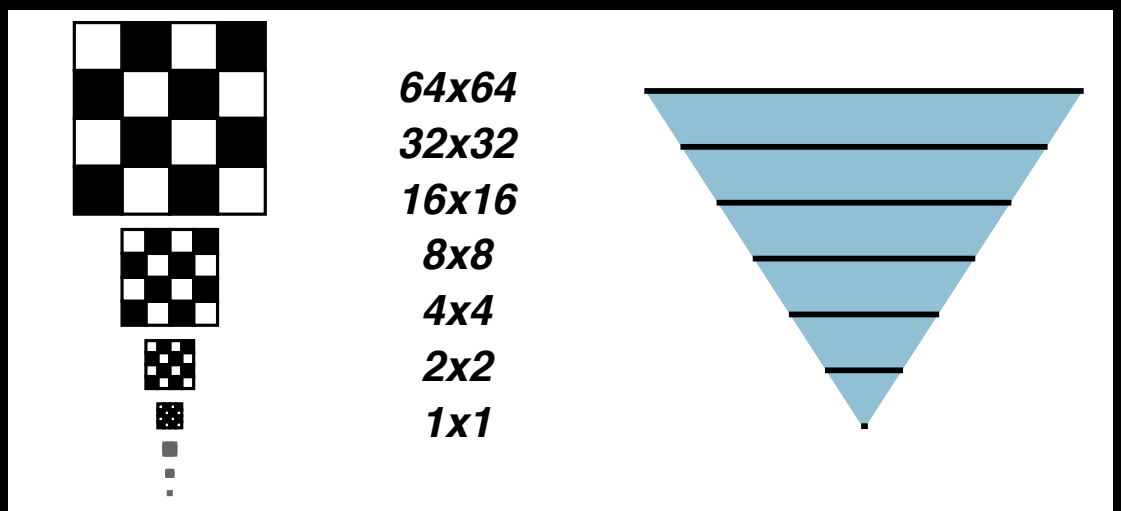

(800) 462-3277
john.lazzaro@gmail.com
Zoom limit in Google Earth has at least 1 meter resolution.
California Memorial Stadium
UC Berkeley's outdoor...
Google
Imagery ©2014 Google, Map data ©2014 Google
Terms
Privacy
Report a problem
20 ft


40 million pixels
john.lazzaro@gmail.com
11 Petabytes for a full-zoom map of the earth's surface
20 million pixels
How do you cache an 11 Petabyte map in hardware, so you can walk and zoom through it at 60 frames/sec? In 1998?
United States
Gulf of Mexico
Cuba
Dominican Republic
Gulf of California
Mexico
Guatemala
Costa Rica

The Clipmap: A Virtual Mipmap

Christopher C. Tanner, Christopher J. Migdal, and Michael T. Jones
Silicon Graphics Computer Systems

Imagine: The same 1 square mile patch of earth, pictured in larger and larger images

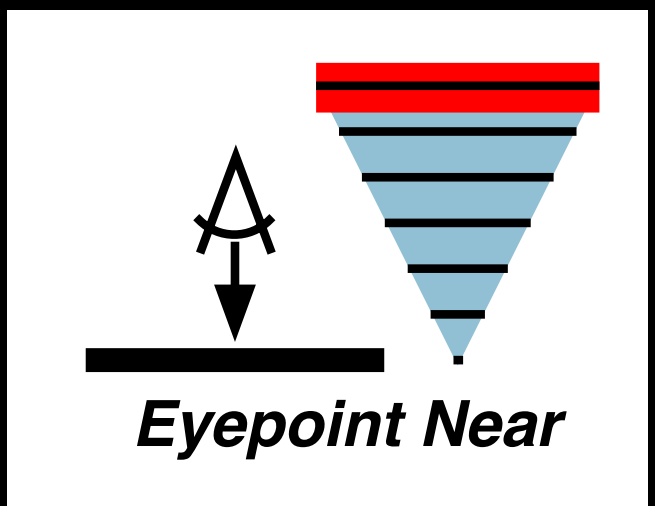


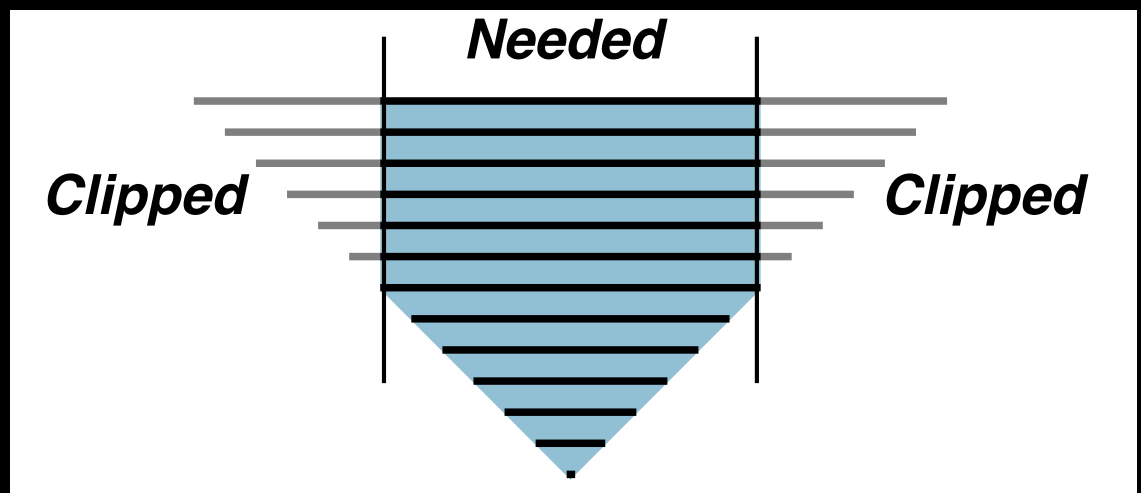
As we go larger and larger, we see more and more details within the 1 square mile … first roads … then cars … then people …

Assume MacBook Air … 1386 x 768 screen …

We are all zoomed in on Google Maps

Top pyramid image is 4K x 4K … Idea: Keep only a 1386 x 768 window of top images in RAM …

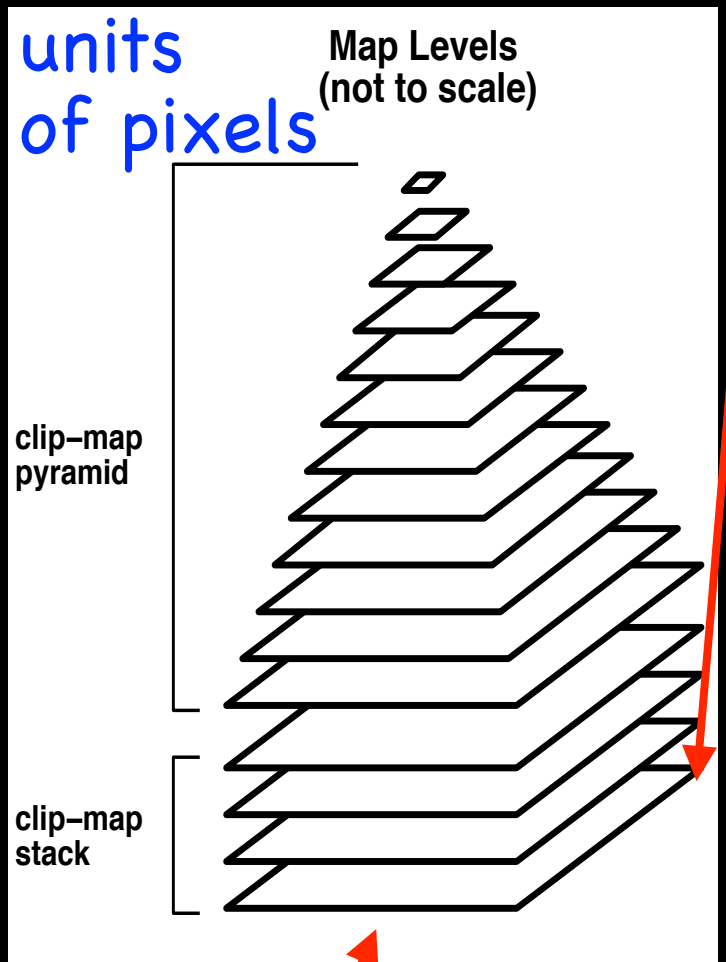




Lets us cache a 1024 x 1024 window of the 11 PB Earth map in 34.7 MB!

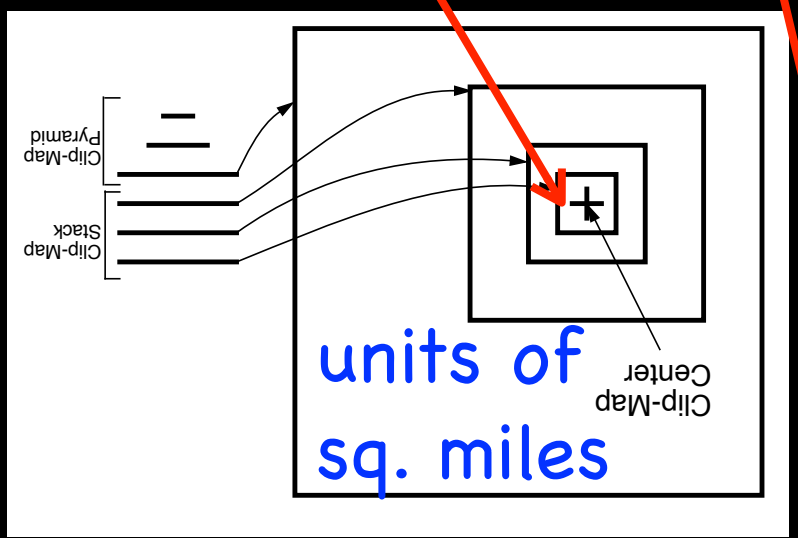
Type and Size	512^2	1024^2	4096^2	32768^2	67108864^2
Full Mipmap	682KB	2.7MB	42.7MB	2.7GB	10923TB
512^2 Clipmap	682KB	1.1MB	2.2MB	3.7MB	9.1MB
1024^2 Clipmap	682KB	2.7MB	6.7MB	12.7MB	34.7MB
2048^2 Clipmap	682KB	2.7MB	18.7MB	42.7MB	131.7MB

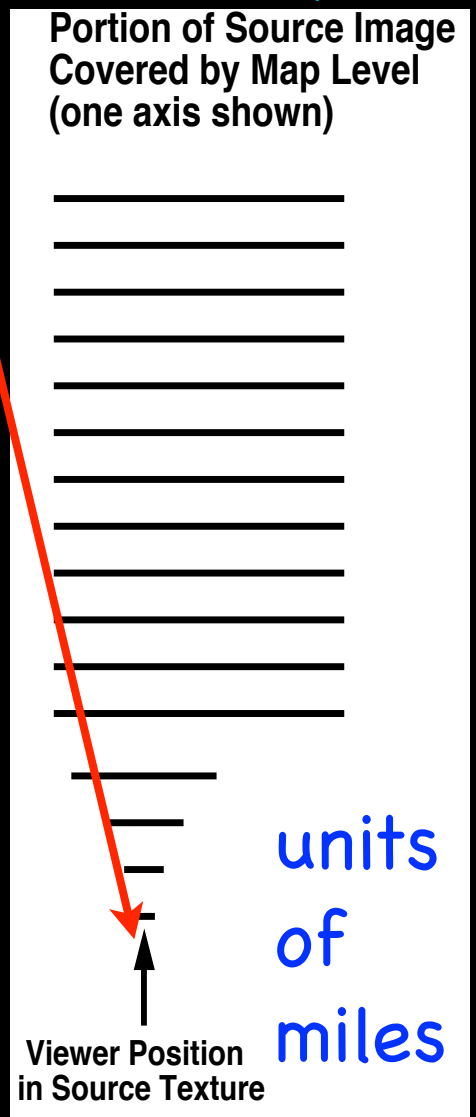
Zoom all the way in ...



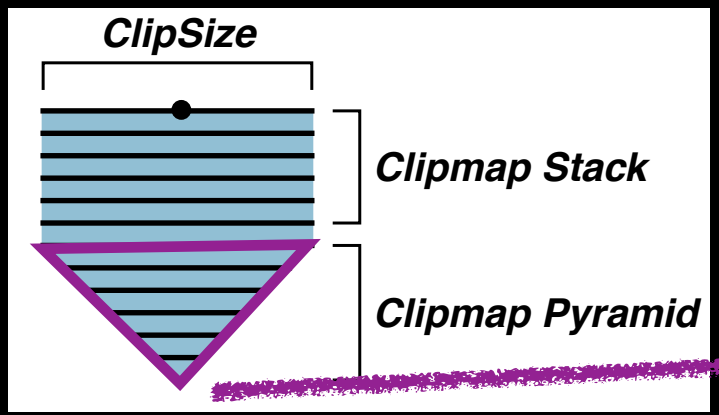
Graphics hardware displays bottom stack image, which fills MacBook Air display.

Bottom stack image shows the smallest part of the 1 mile sq. patch of the Earth of any stack image.



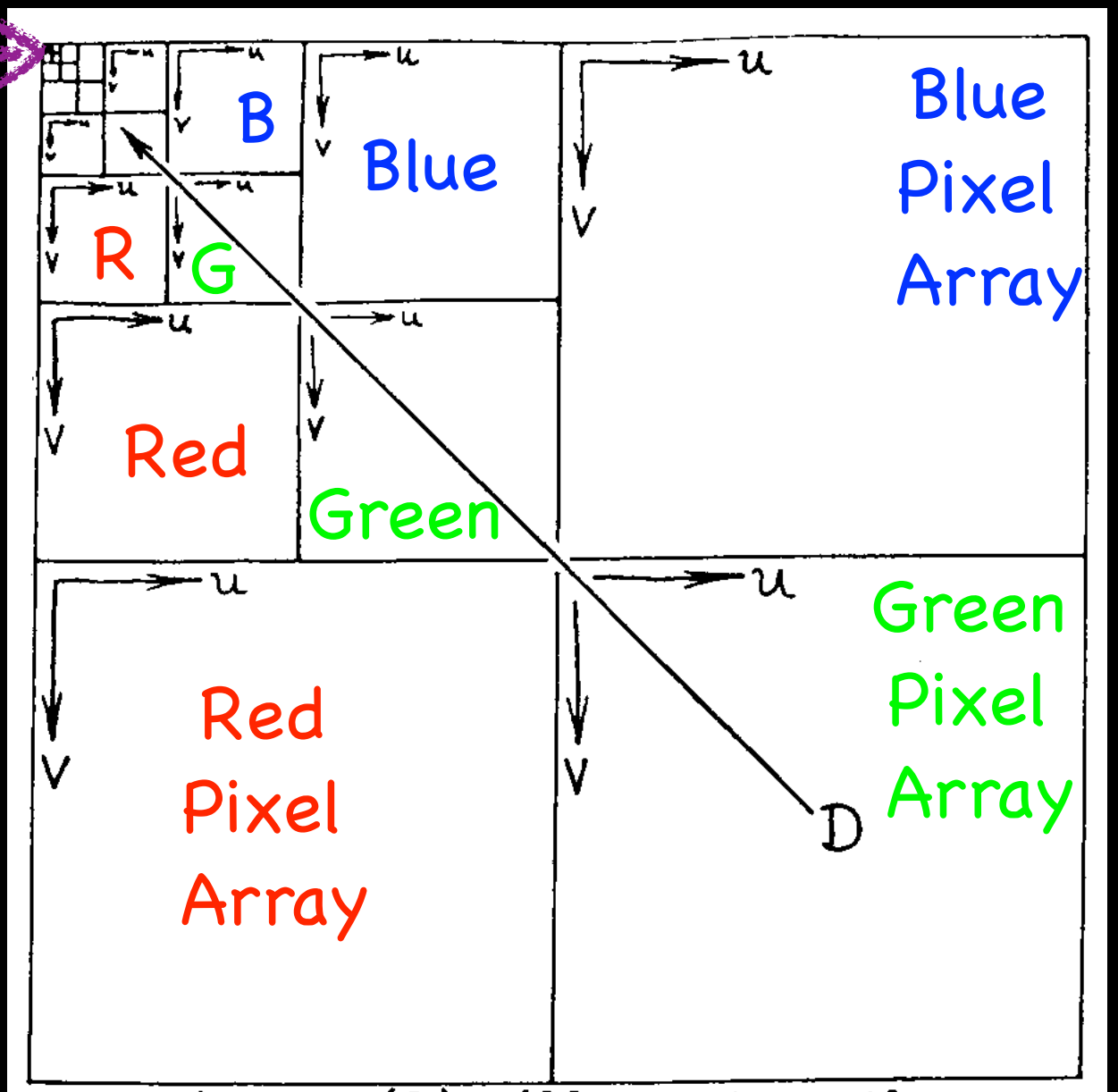


Hardware interpolation of stack levels.



Efficient memory layout of pyramidal part of ClipMap

^ This arrow points to tip of pyramid.



Updating the image as we move over the Earth

Toroidal memory indexing. Never move pixels !

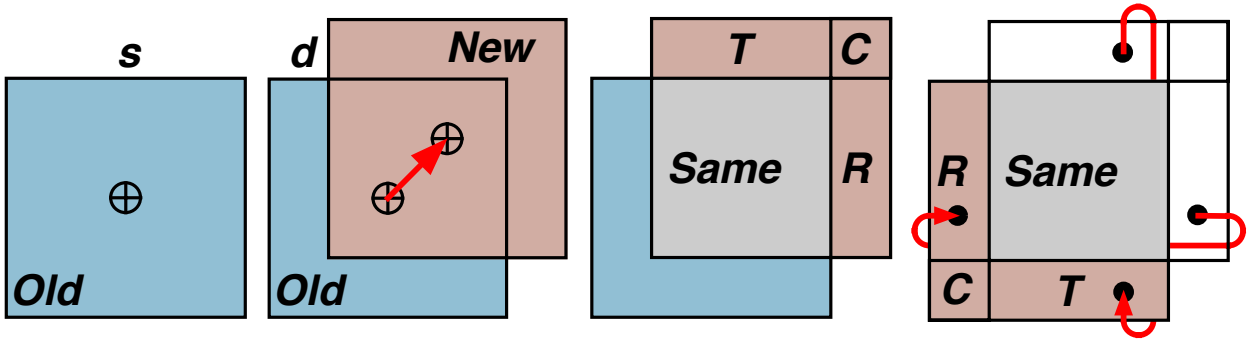


Figure 7: 2D Image Roam using Toroidal Addressing

When we move over the earth at a certain depth ... all depths are toroidally updated.

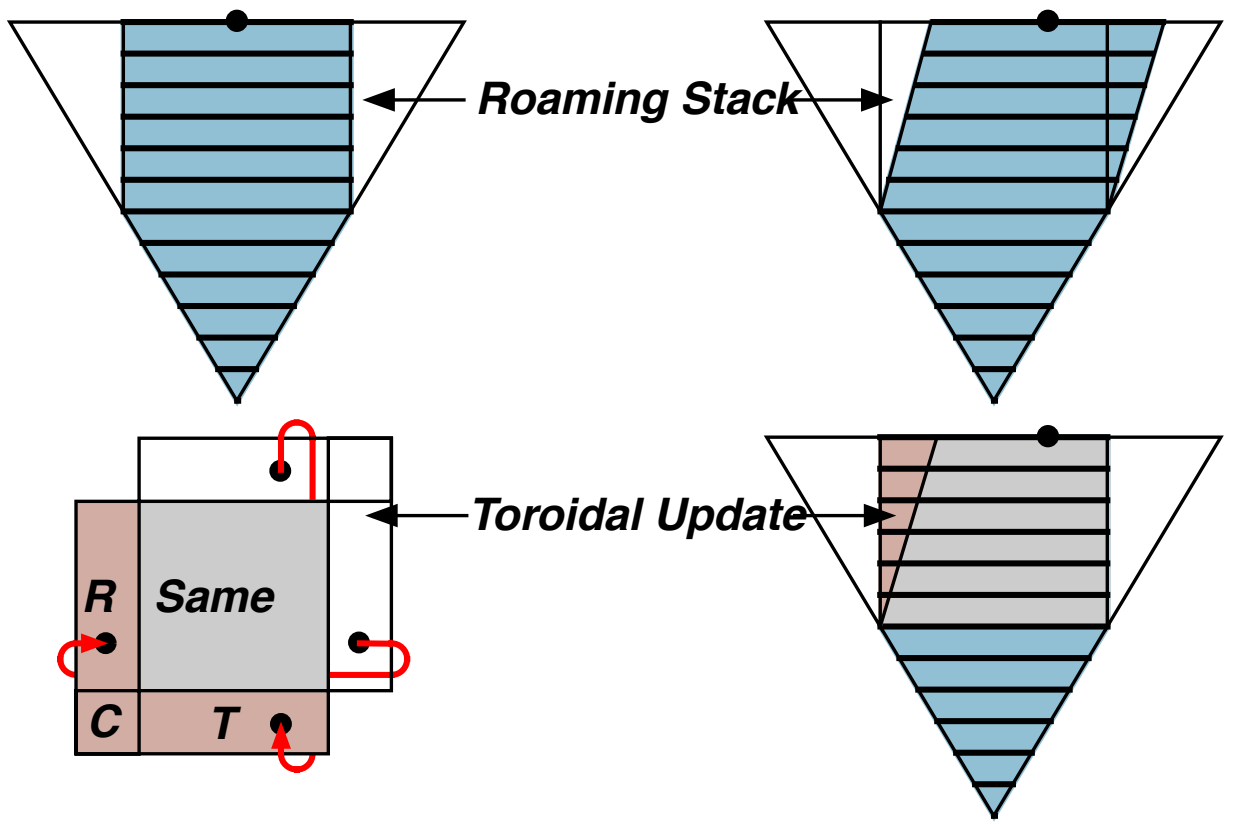


Figure 8: 2D Image Roam of Complete Stack

Virtual ClipMaps

16-level
Real ClipMap
in hardware ..
27-level
Virtual Clipmap
managed by
driver software.

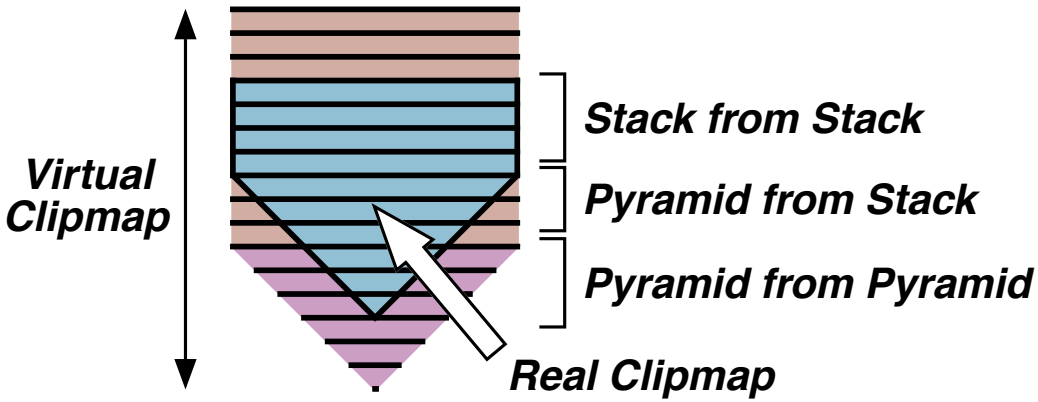


Figure 9: Virtual Addressing in Stack and Pyramid Levels

"Disk" became Google Maps "cloud"

Driver software
also caches
larger regions in
main memory,
disk paging.

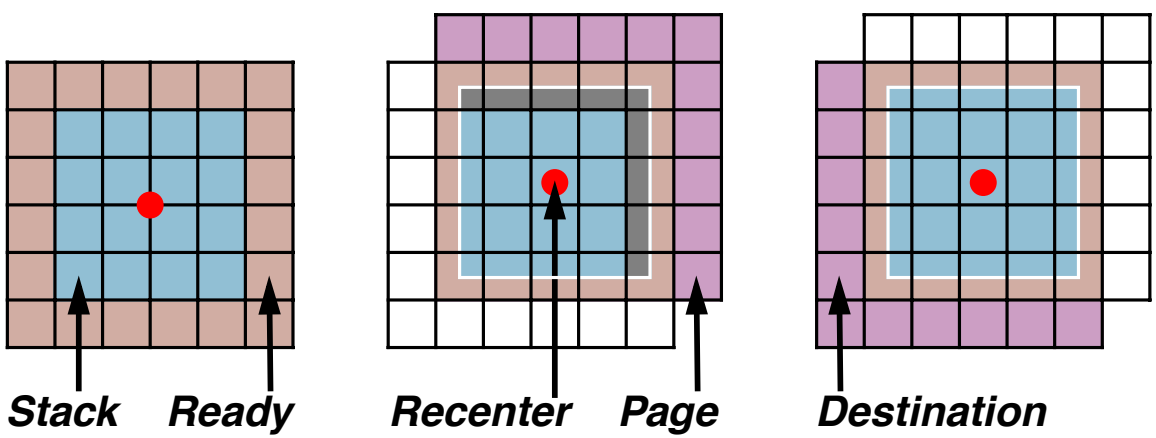


Figure 10: Tiled Main-Memory Image Cache

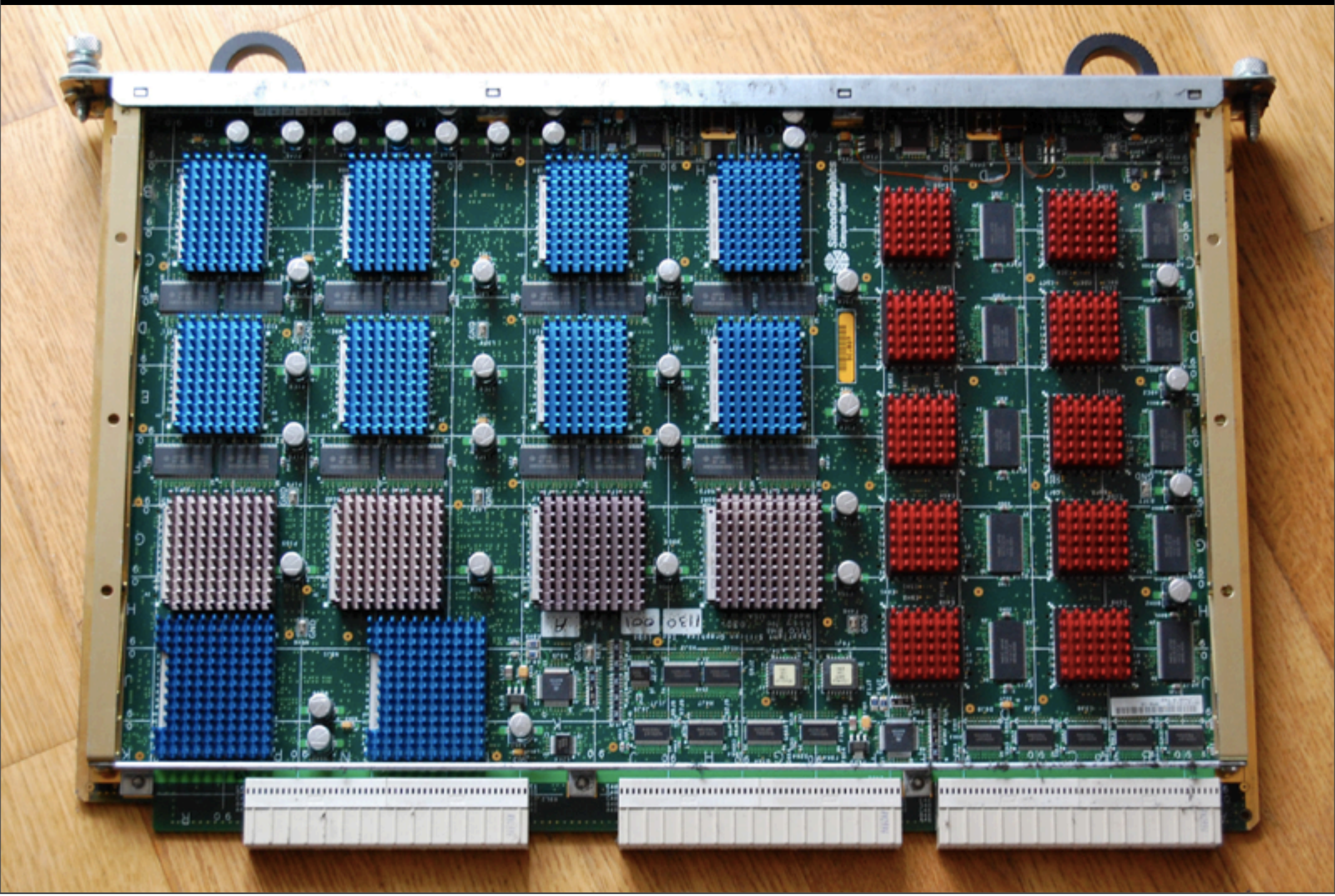
SGI Onyx2 (1998)

6 board Infinite Reality graphics (per pipeline)

Many copies of 12 custom ASICS ... full configuration had 251 M transistors

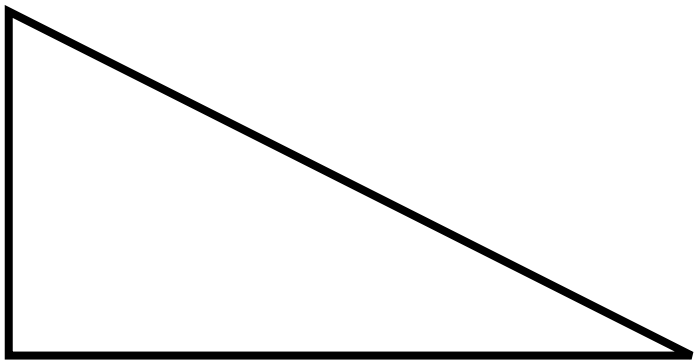
$1M+ ...
time machine for single-chip GPUs.

Raster Manager board … 4 different ASICs



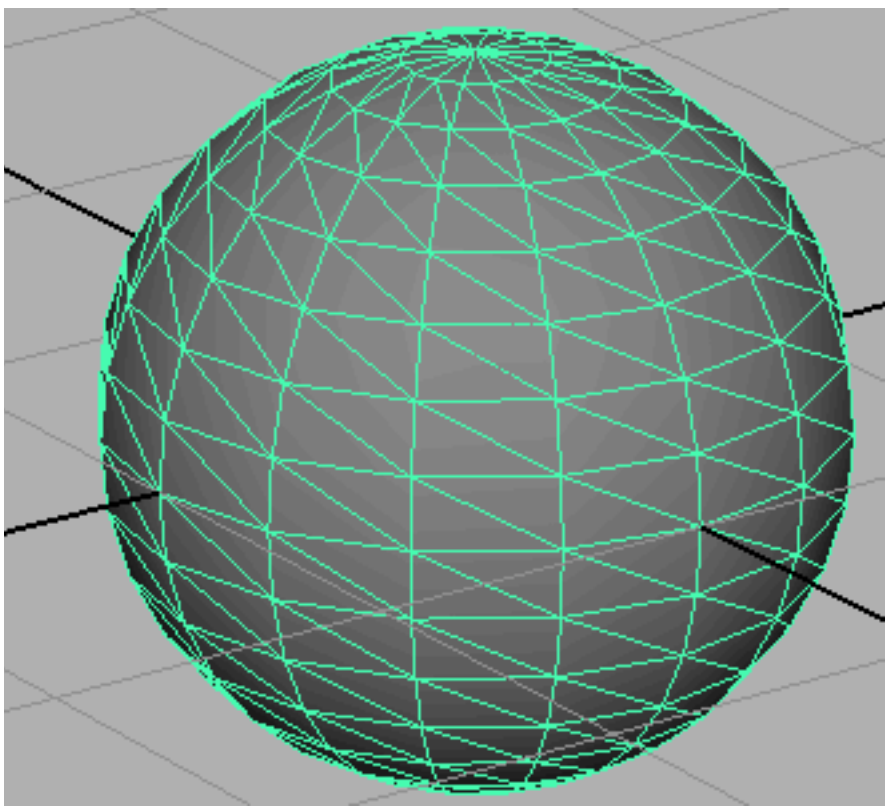

62-3277
john.lazzaro@gmail.com
How does an Earth ClipMap become a globe?
United States
Dominican Republic
Cuba
Gulf of Mexico
Gulf of California
Mexico
Guatemala
Costa Rica

From a triangle ... to a globe ...



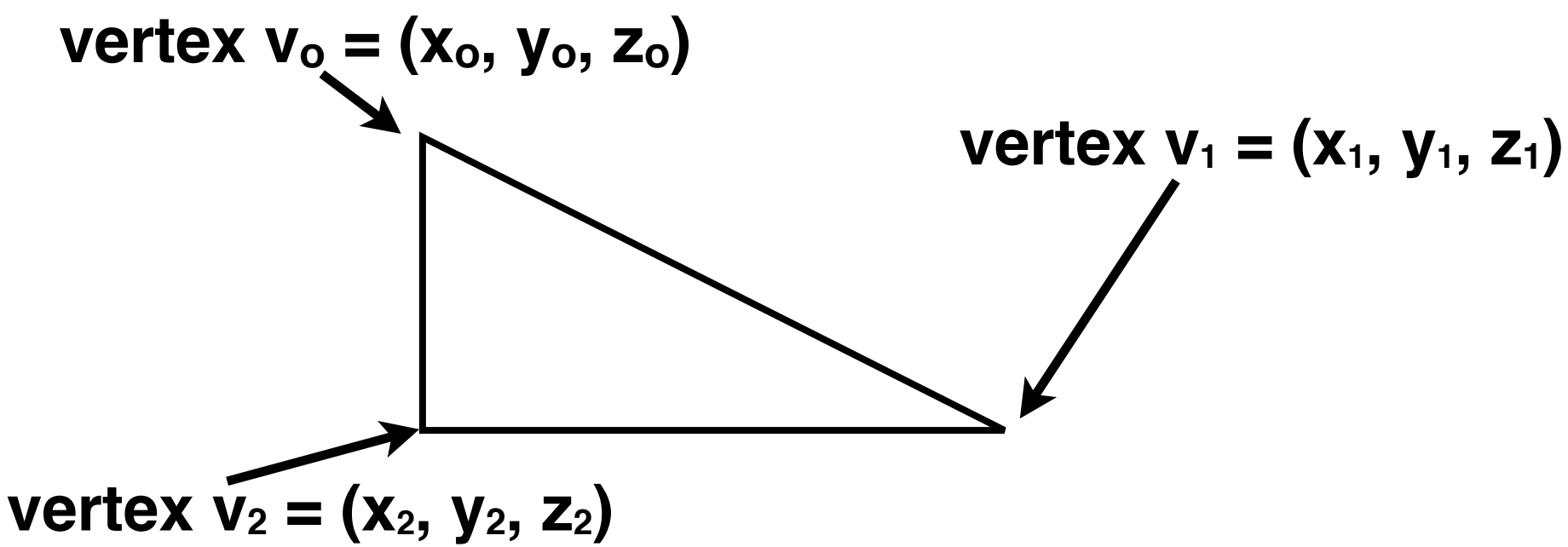
Simplest closed shape that may be defined by straight edges.

A sphere whose faces are made up of triangles. With enough triangles, the curvature of the sphere can be made arbitrarily smooth.

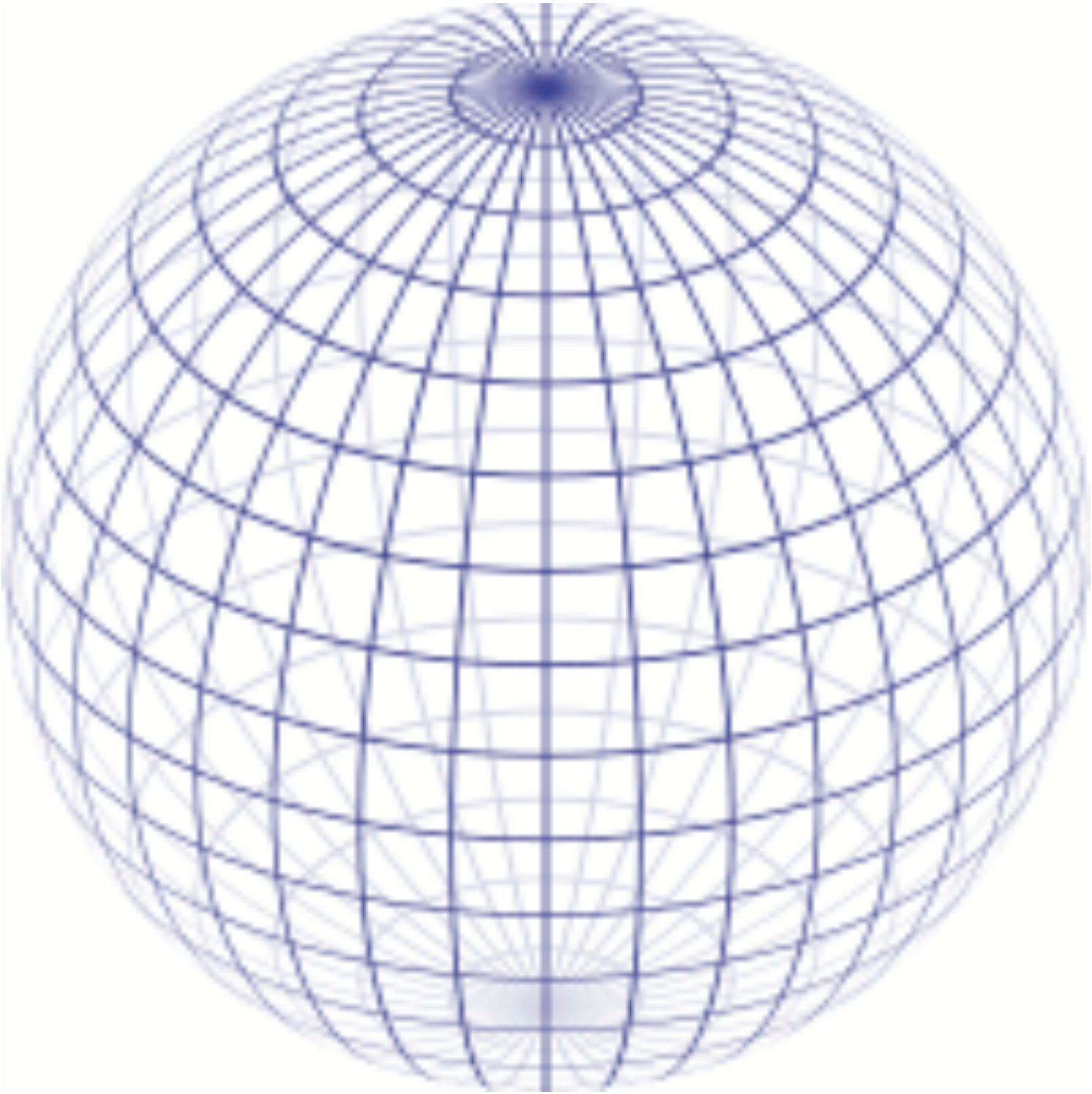


Triangle defined by 3 vertices

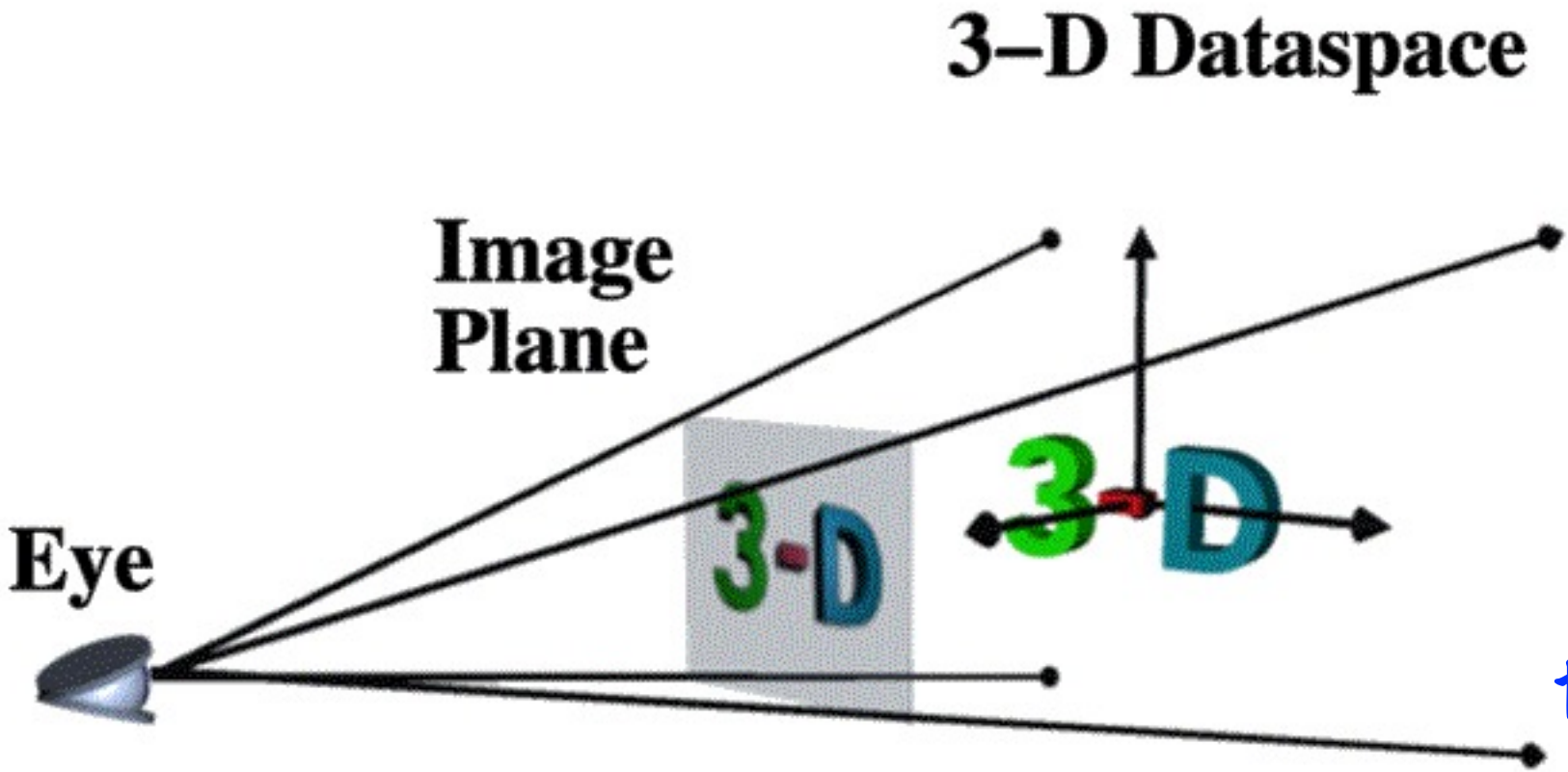
By transforming ($v' = f(v)$) all vertices in a 3-D object (like a globe), you can move it in the 3-D world, change it's size, rotate it, etc.



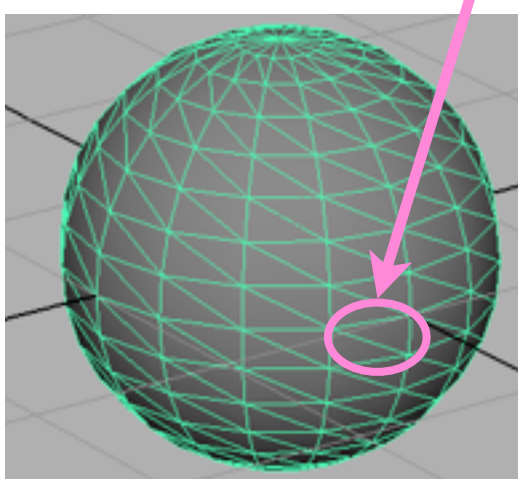
If a globe has 10,000 triangles, need to transform 30,000 vertices to move it in a 3-D scene ... per frame!



We see a 2-D window into the 3-D world



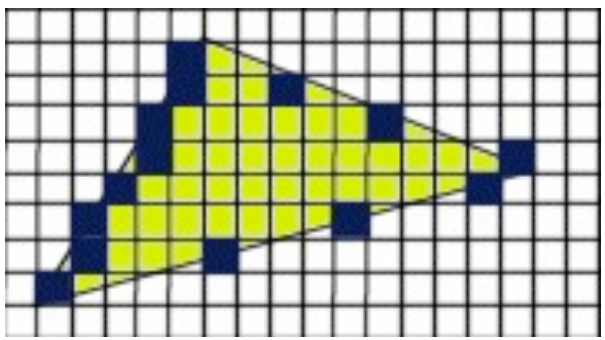
Let's follow one 3-D triangle.



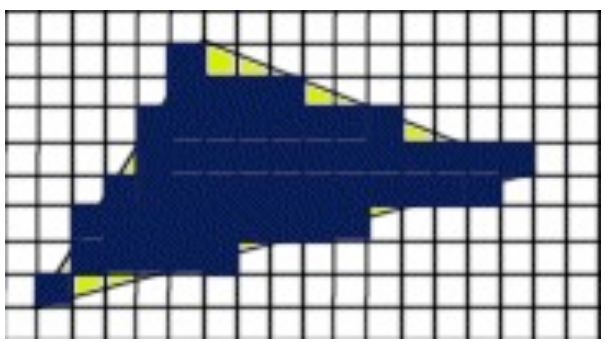
From 3-d triangles to screen pixels

First, project each 3-D triangle that might "face" the "eye" onto the image plane.

Then, create "pixel fragments" on the boundary of the image plane triangle



Then, create "pixel fragments" to fill in the triangle (rasterization).

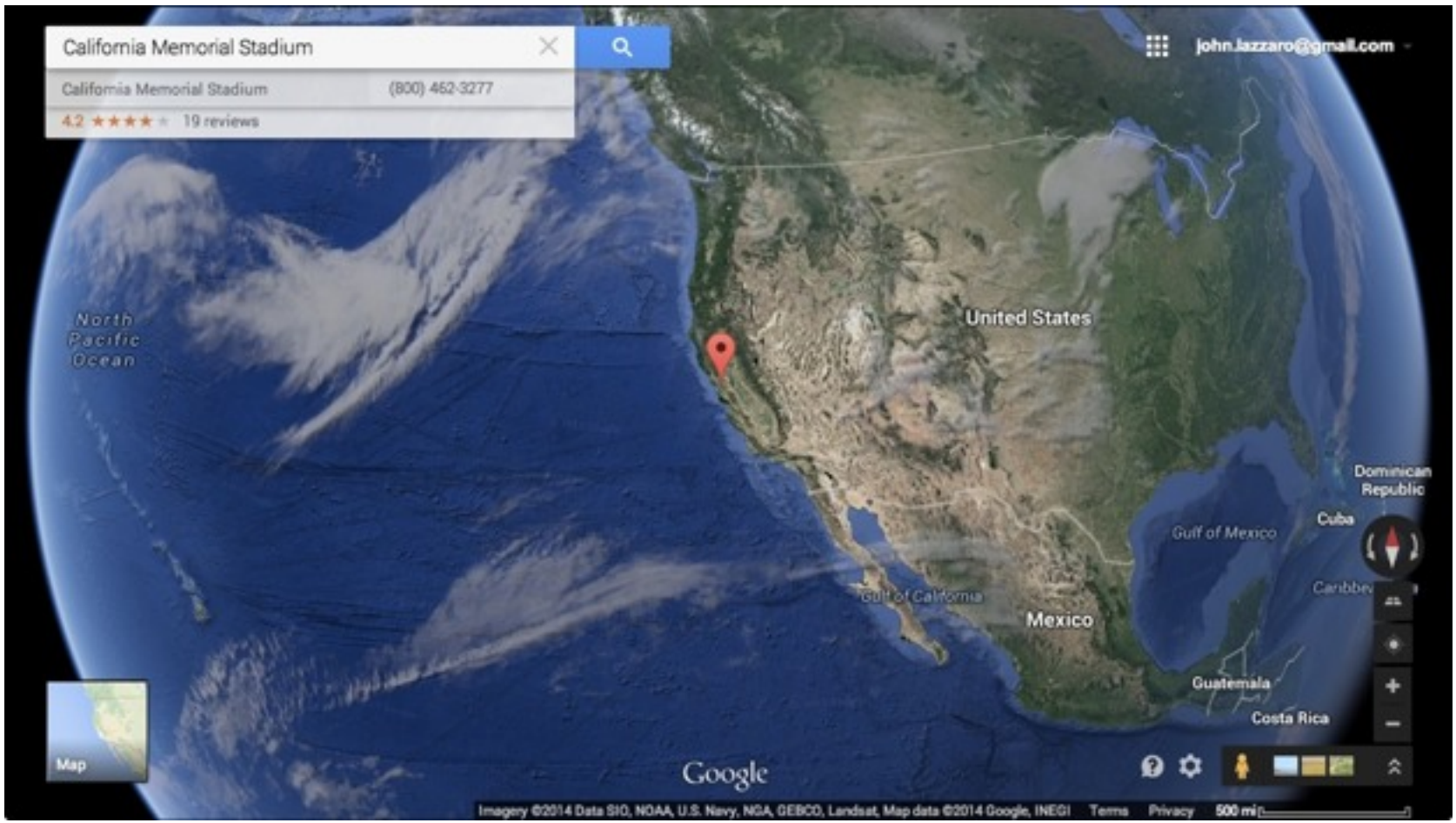


Why "pixel fragments"? A screen pixel color might depend on many triangles (example: a glass globe).

Process each fragment to "shade" it.

11 PB Earth Map: Lives in Google cloud. Cached locally in a ClipMap. We "map" the correct Earth "texture" onto each pixel fragment during shading.

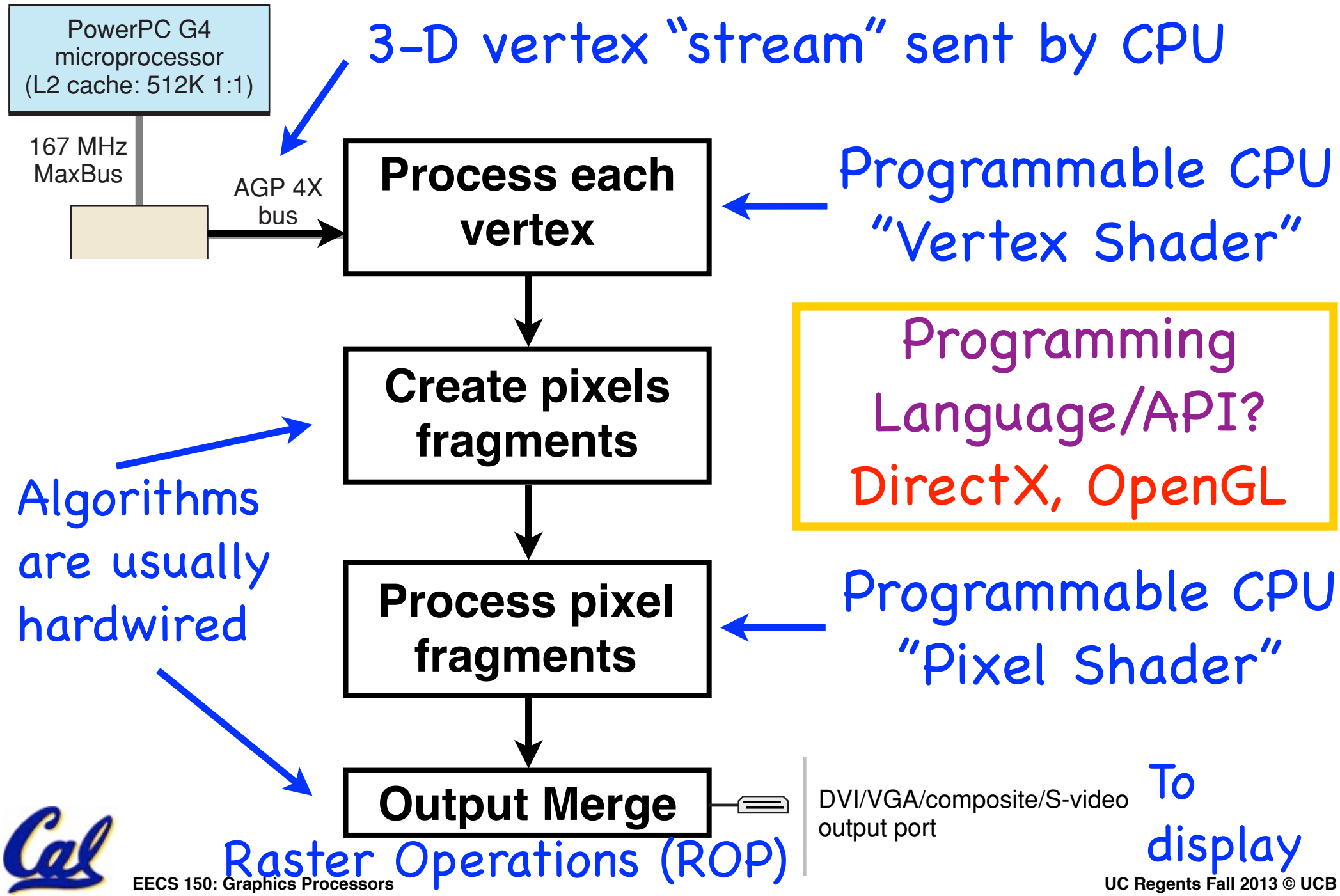
Final step: Output Merge. Assemble pixel fragments to make final 2-d image pixels.



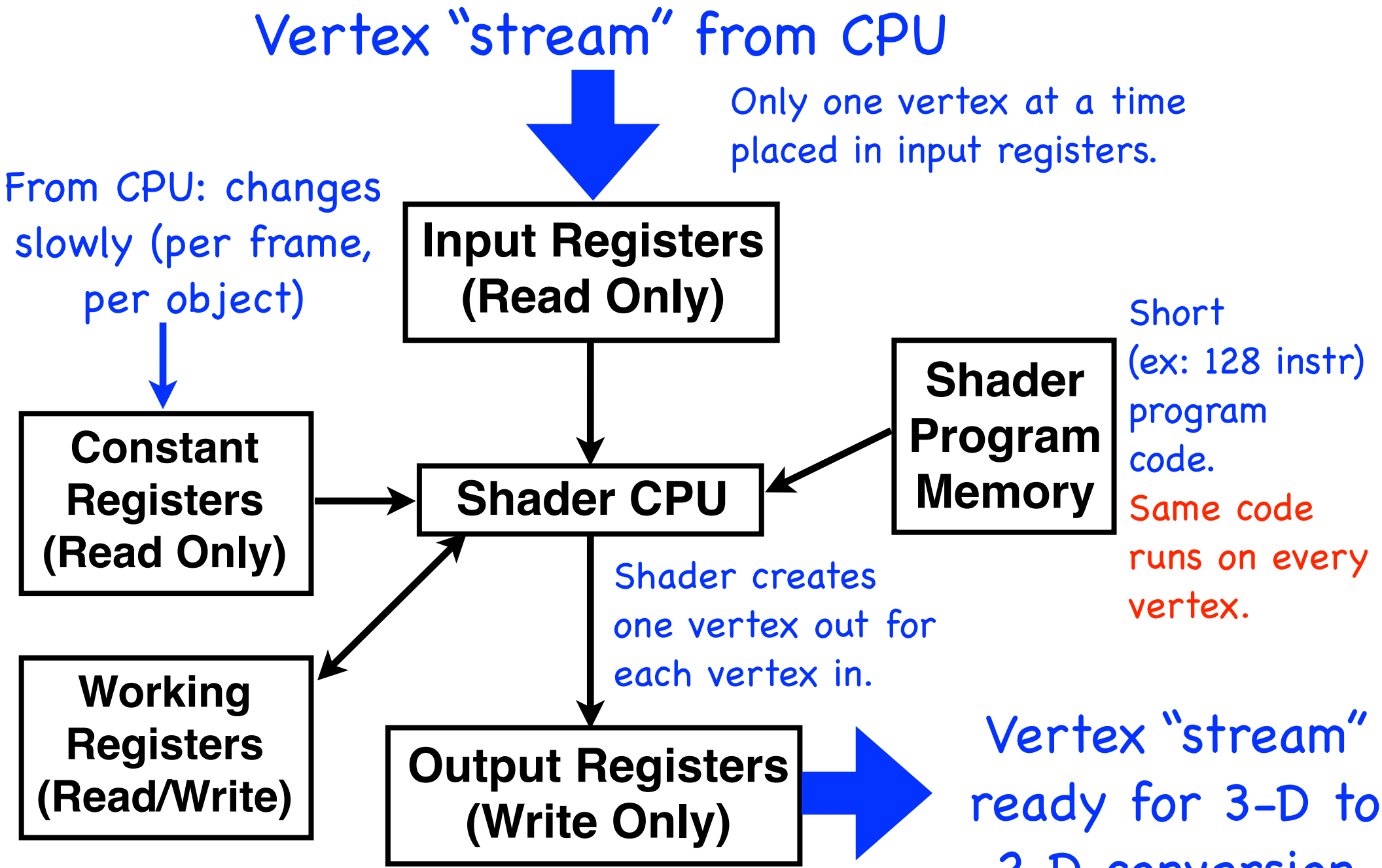
Graphics Acceleration

Next: Back to architecture ...

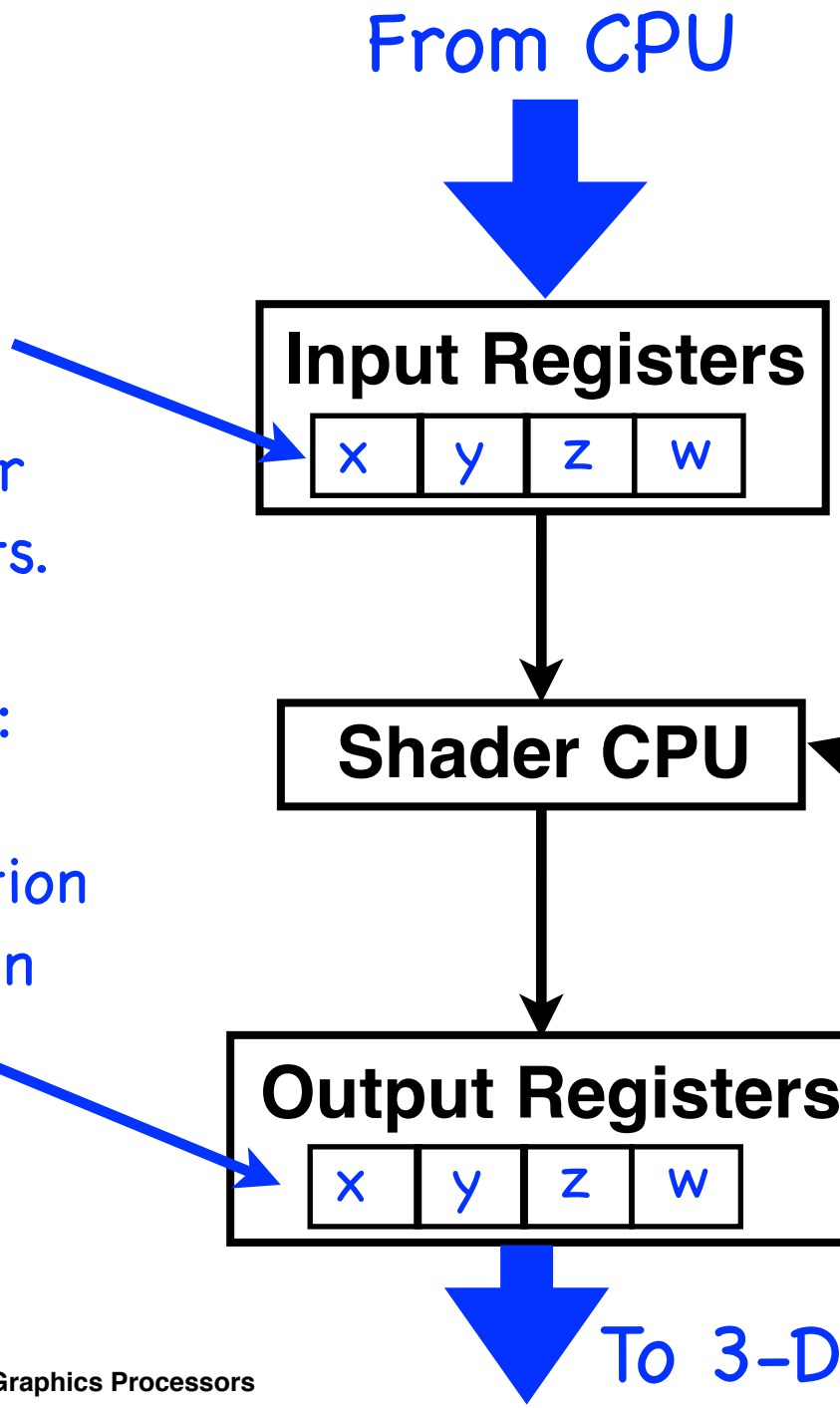
The graphics pipeline in hardware (2004)



Vertex Shader: A "stream processor"



Optimized instructions and data formats



128-bit registers, holding four 32-bit floats.

Typical use: (x,y,z,w) representation of a point in 3-D space.

Typical instruction:

rsq dest src

dest.{x,y,z,w} = 1.0/sqrt(abs(src.w)).

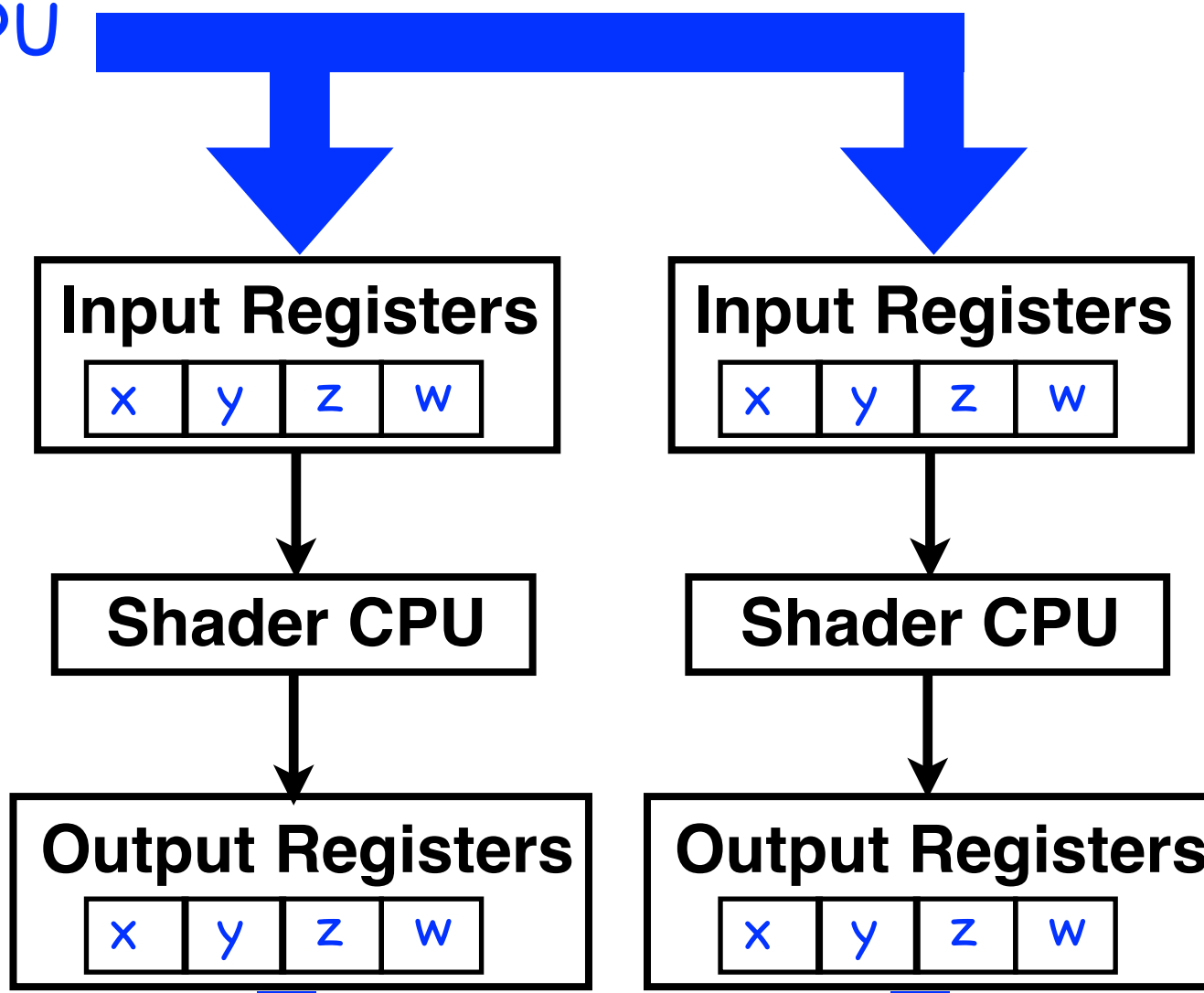
If src.w=0, dest ∞.

Shader Program Memory

The 1/sqrt() function is often used in graphics.

Easy to parallelize: Vertices independent

From CPU



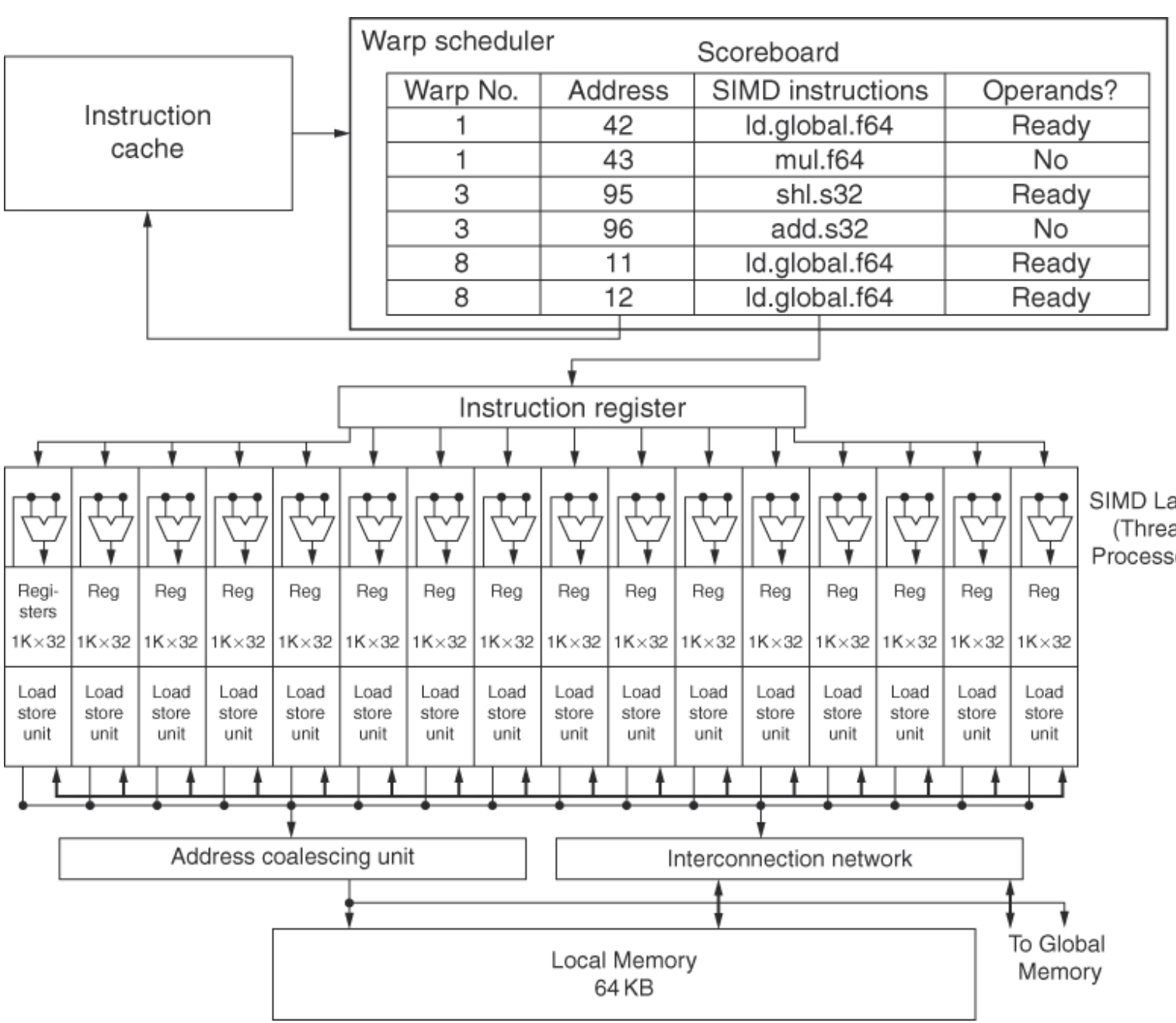
Caveat: Care might be needed when merging streams.

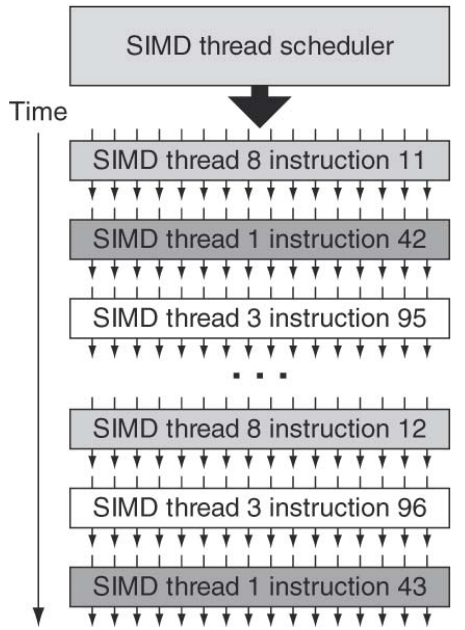
Why? 3-D to 2-D may expect triangle vertices in order.

Shader CPUs easy to multithread. Also, SIMD-like control.

To 3-D/ 2-D

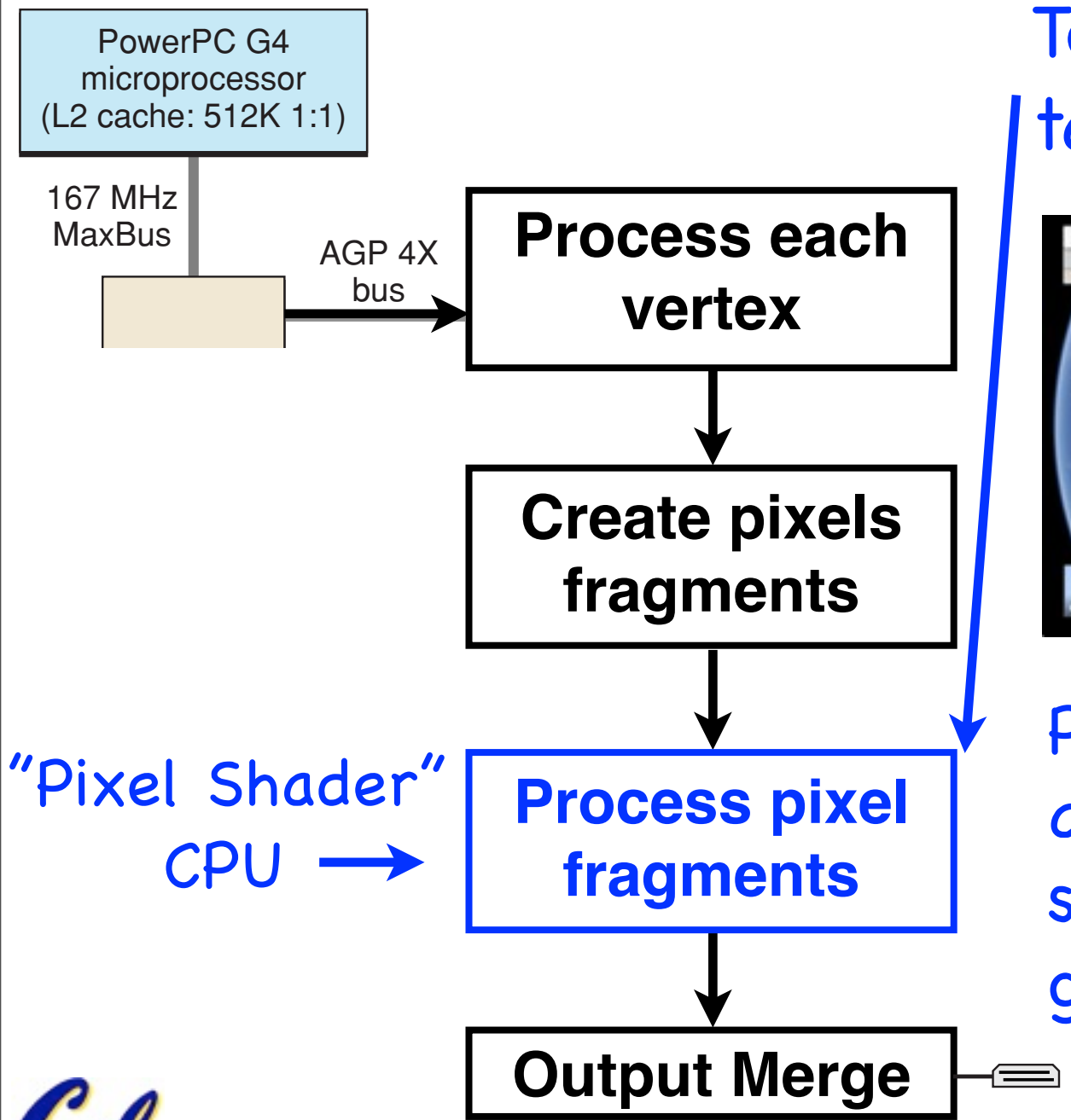
Multithreading? SIMD? Recall last lecture!



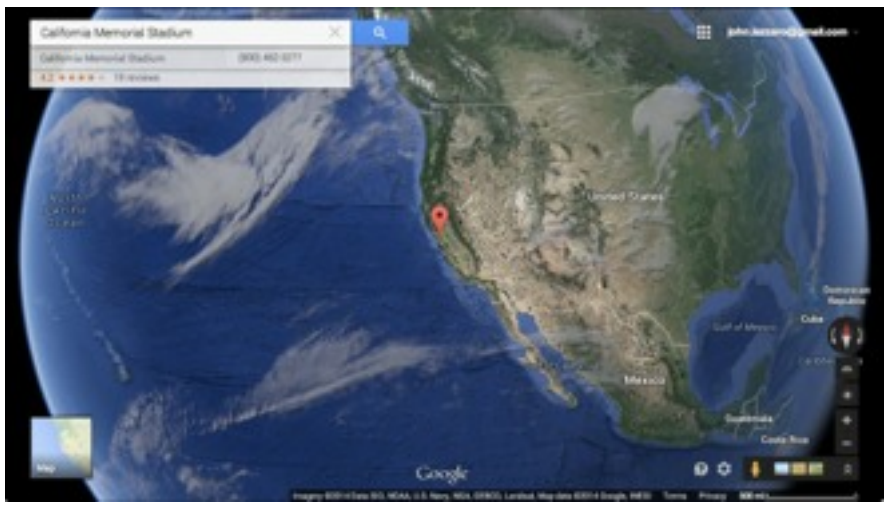


From Hennessy and Patterson textbook.

Pixel shader specializations ...



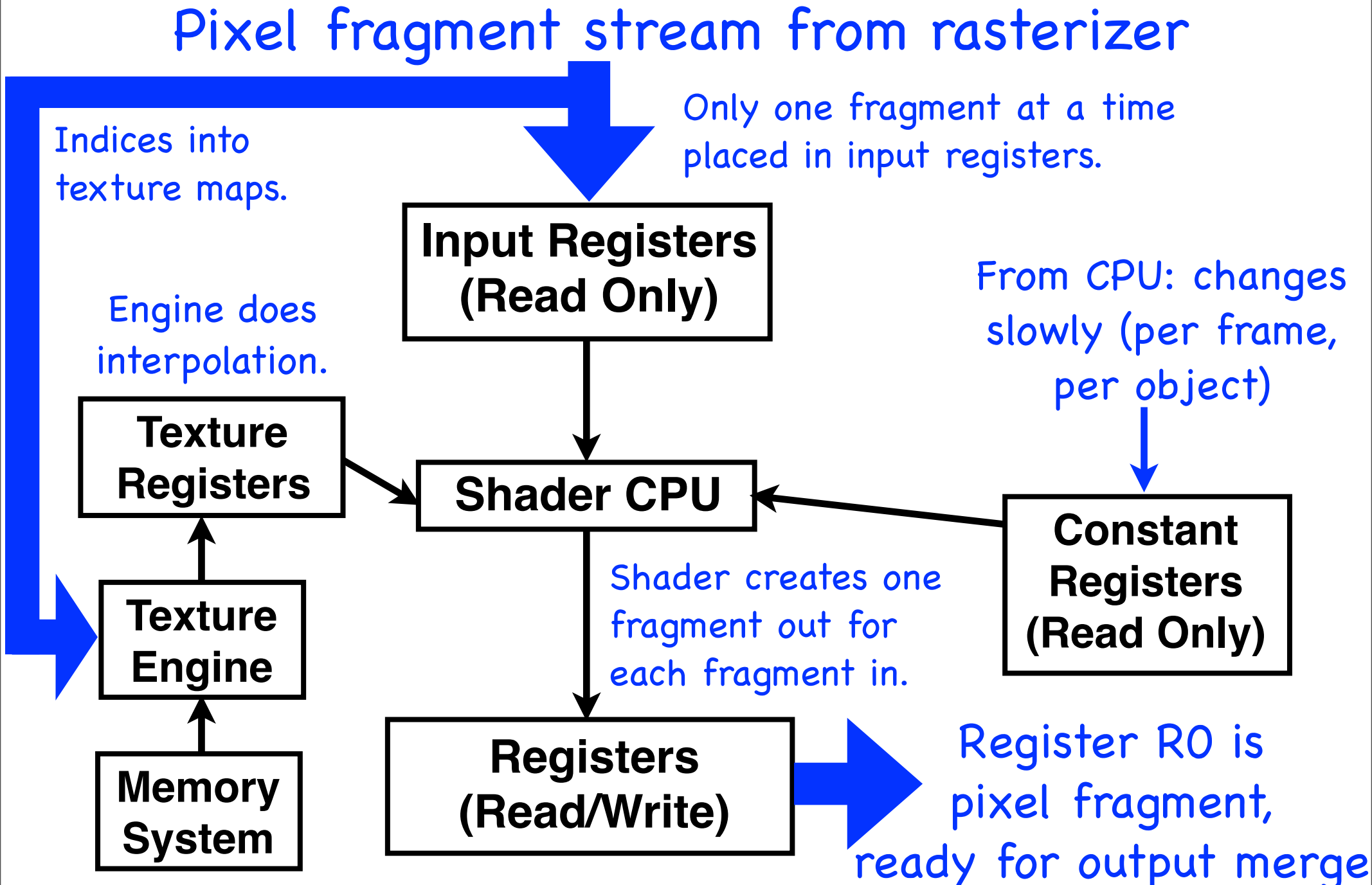
Texture maps (look-up tables) play a key role.



Pixel shader needs fast access to ClipMap to shade globe (via graphics card RAM).

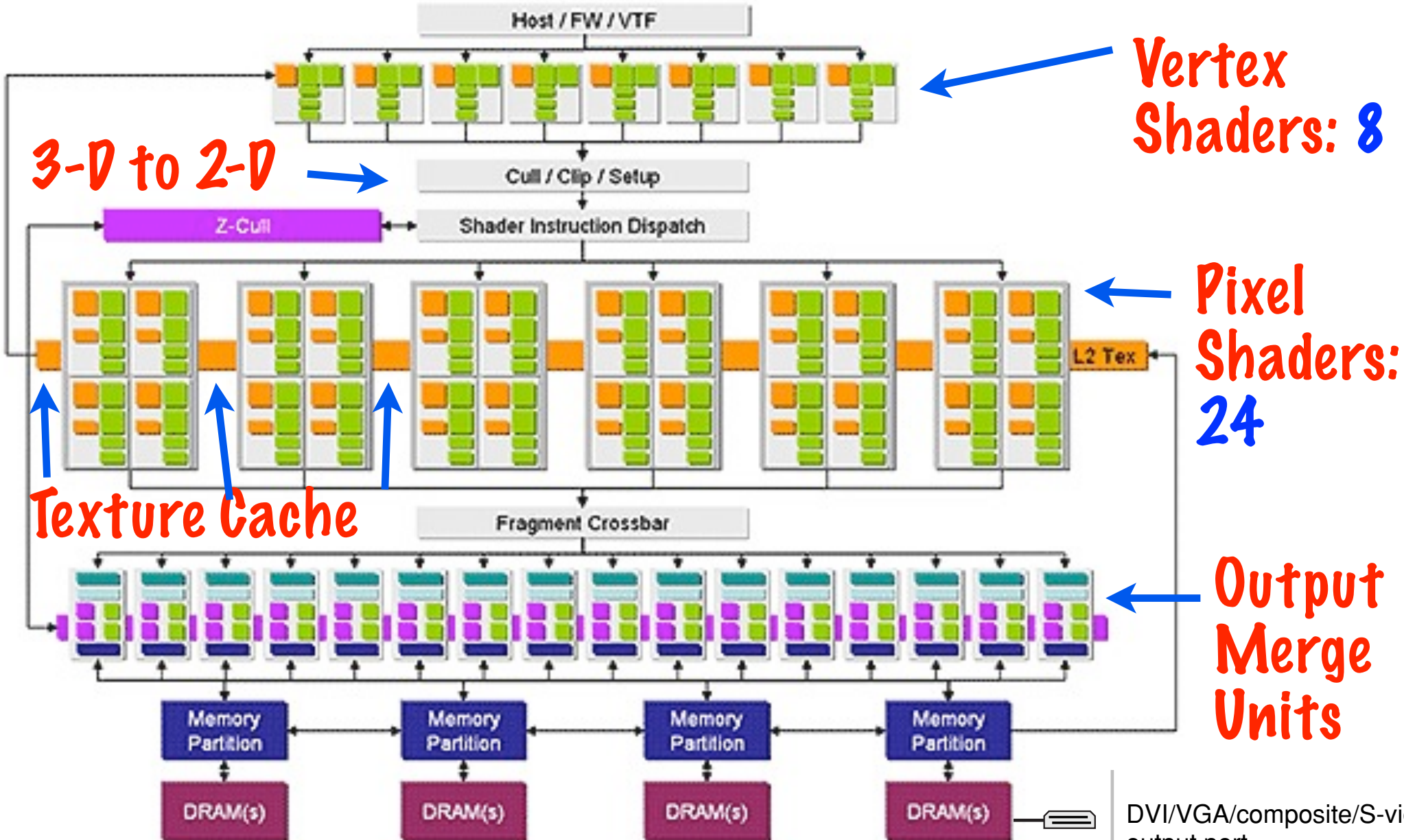
DVI/VGA/composite/S-video output port

Pixel Shader: Stream processor + Memory



Example (2006): Nvidia GeForce 7900

278 Million Transistors, 650 MHz clock, 90 nm process



Break Time ...

Play

Next: Unified architectures

Basic idea: Replace specialized logic (vertex shader, pixel shader, hardwired algorithms) with many copies of one unified CPU design.

Unified Architectures

Consequence: You no longer "see" the graphics pipeline when you look at the architecture block diagram.

Designed for: DirectX 10 (Microsoft Vista), and new non-graphics markets for GPUs.

DirectX 10 (Vista): Towards Shader Unity

Earlier APIs: Pixel and Vertex CPUs very different …

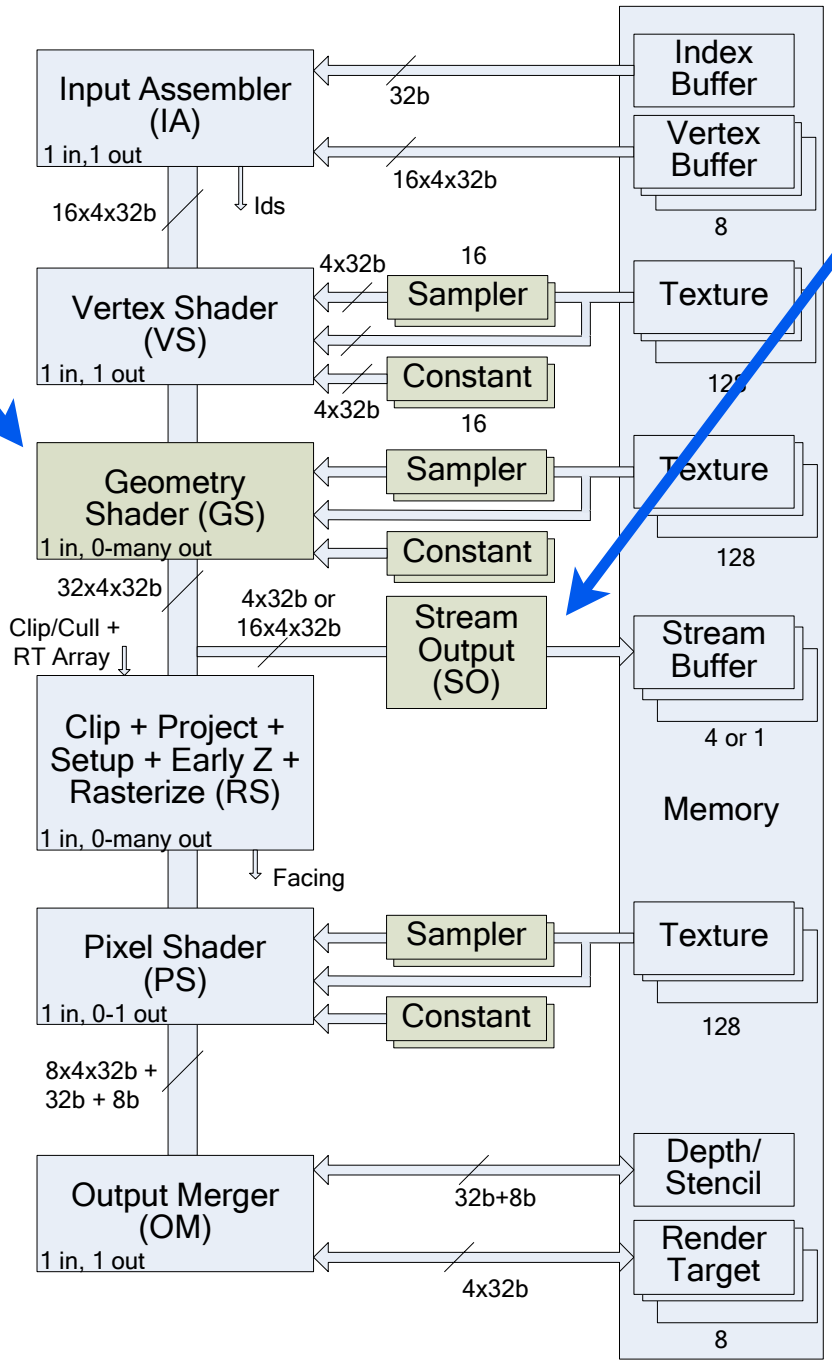
Feature	1.1 2001	2.0 2002	3.0 2004†	4.0 2006
instruction slots	128	256	≥512	≥64K
	4+8‡	32+64‡	≥512	
constant registers	≥96	≥256	≥256	16x4096
	8	32	224	
tmp registers	12	12	32	4096
	2	12	32	
input registers	16	16	16	16
	4+2§	8+2§	10	32
render targets	1	4	4	8
samplers	8	16	16	16
textures			4	128
	8	16	16	
2D tex size			2Kx2K	8Kx8K
integer ops				✓
load op				✓
sample offsets				✓
transcendental ops	✓	✓	✓	✓
		✓	✓	
derivative op			✓	✓
flow control		static	stat/dyn	dynamic
			stat/dyn	

Table 1: Shader model feature comparison summary.

DirectX 10: Many specs are identical for Pixel and Vertex CPUs

DirectX 10 : New Pipeline Features ...

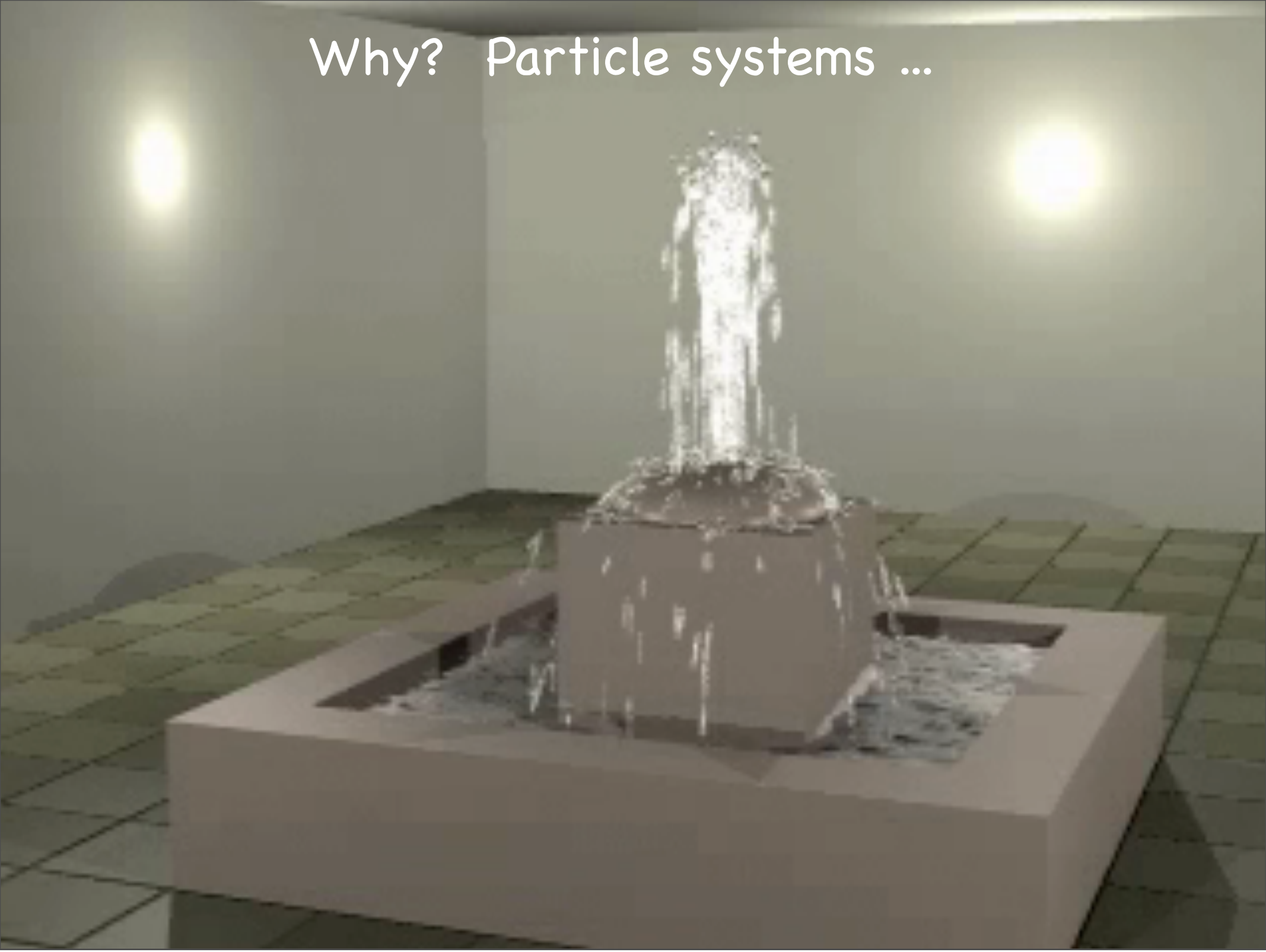
Geometry Shader: Lets a shader program create new triangles.



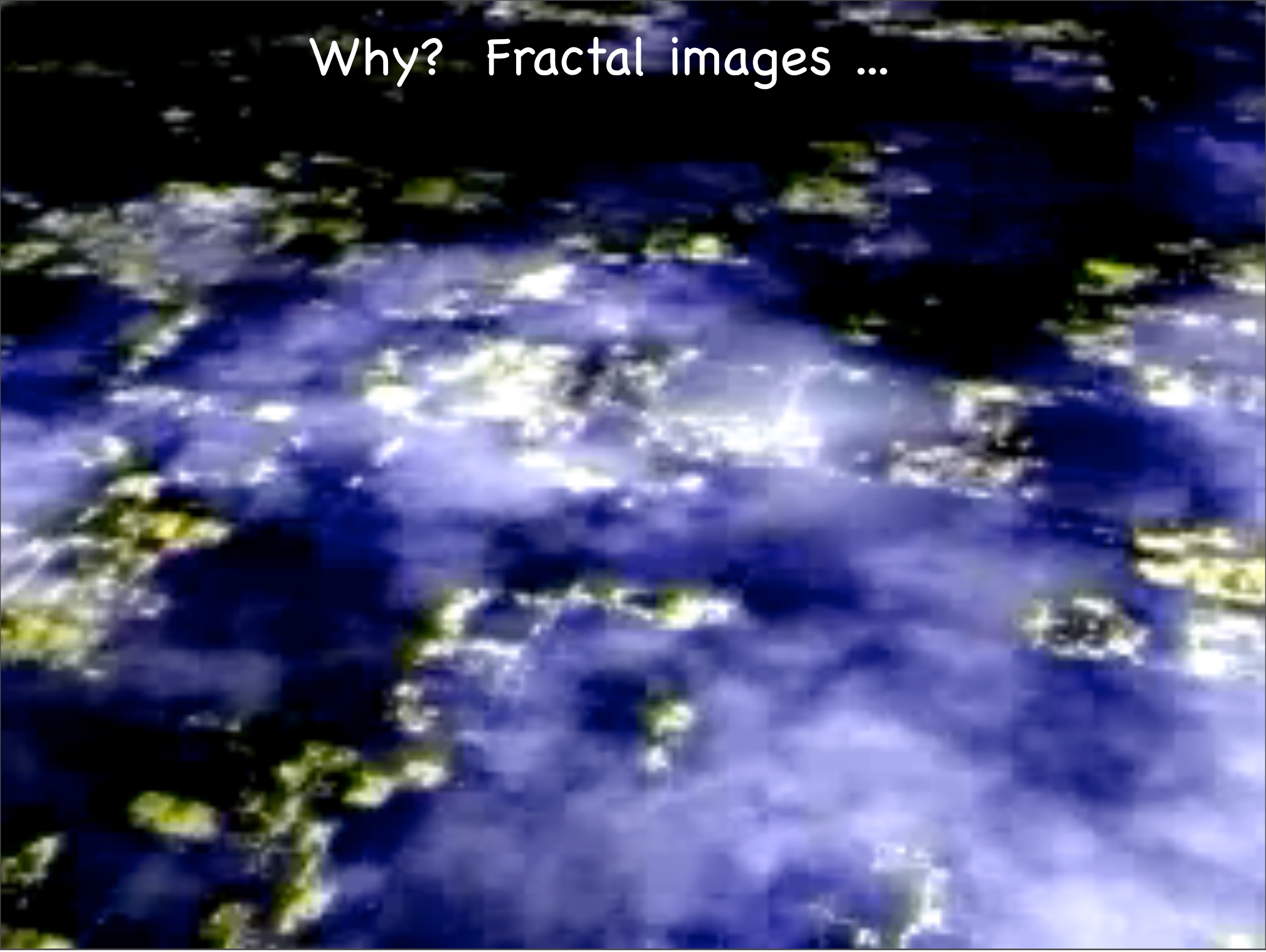
Stream Output: Lets vertex stream recirculate through shaders many times ... (and also, back to CPU)

Also: Shader CPUs are more like RISC machines in many ways.

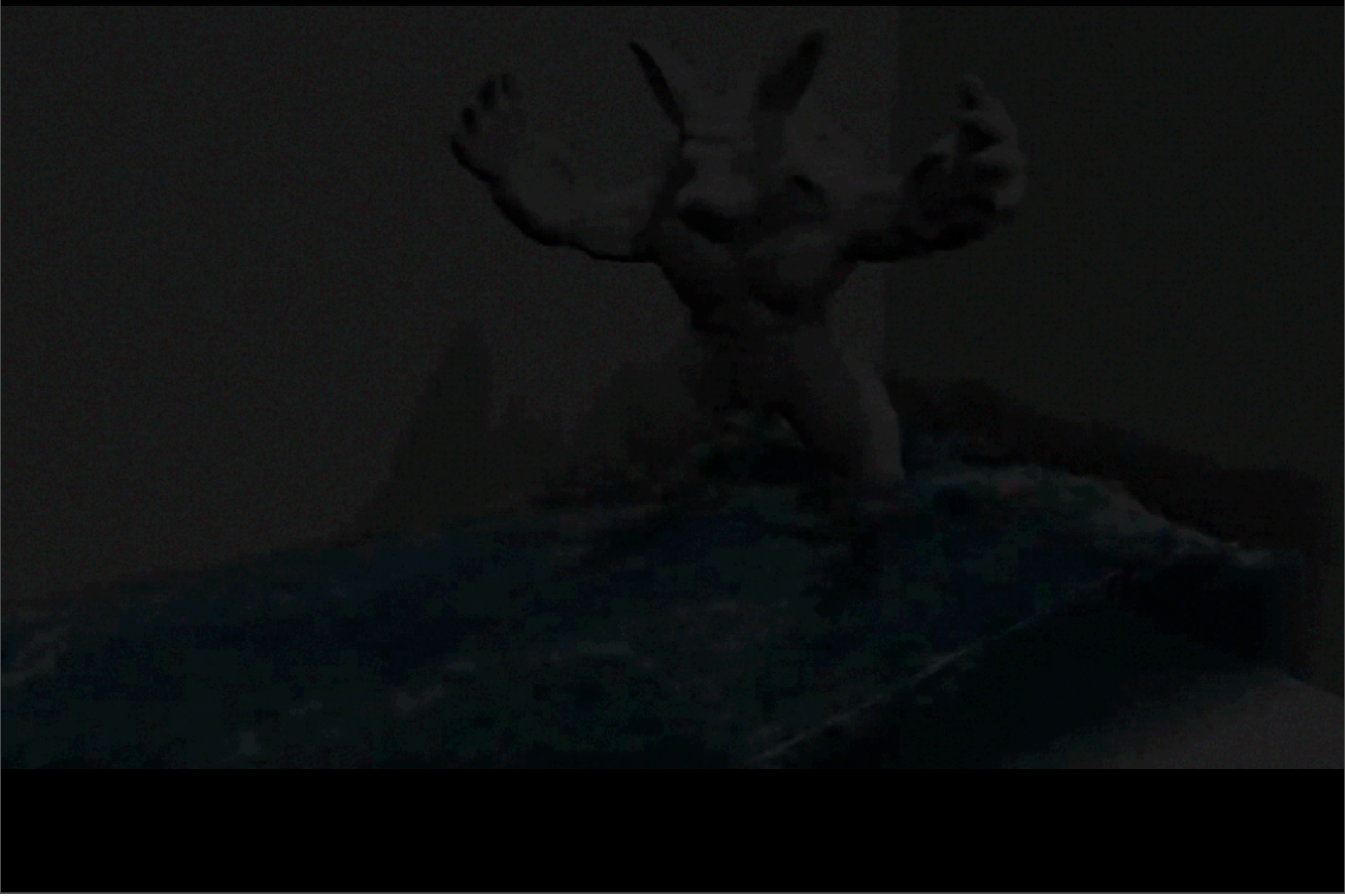
Why? Particle systems ...



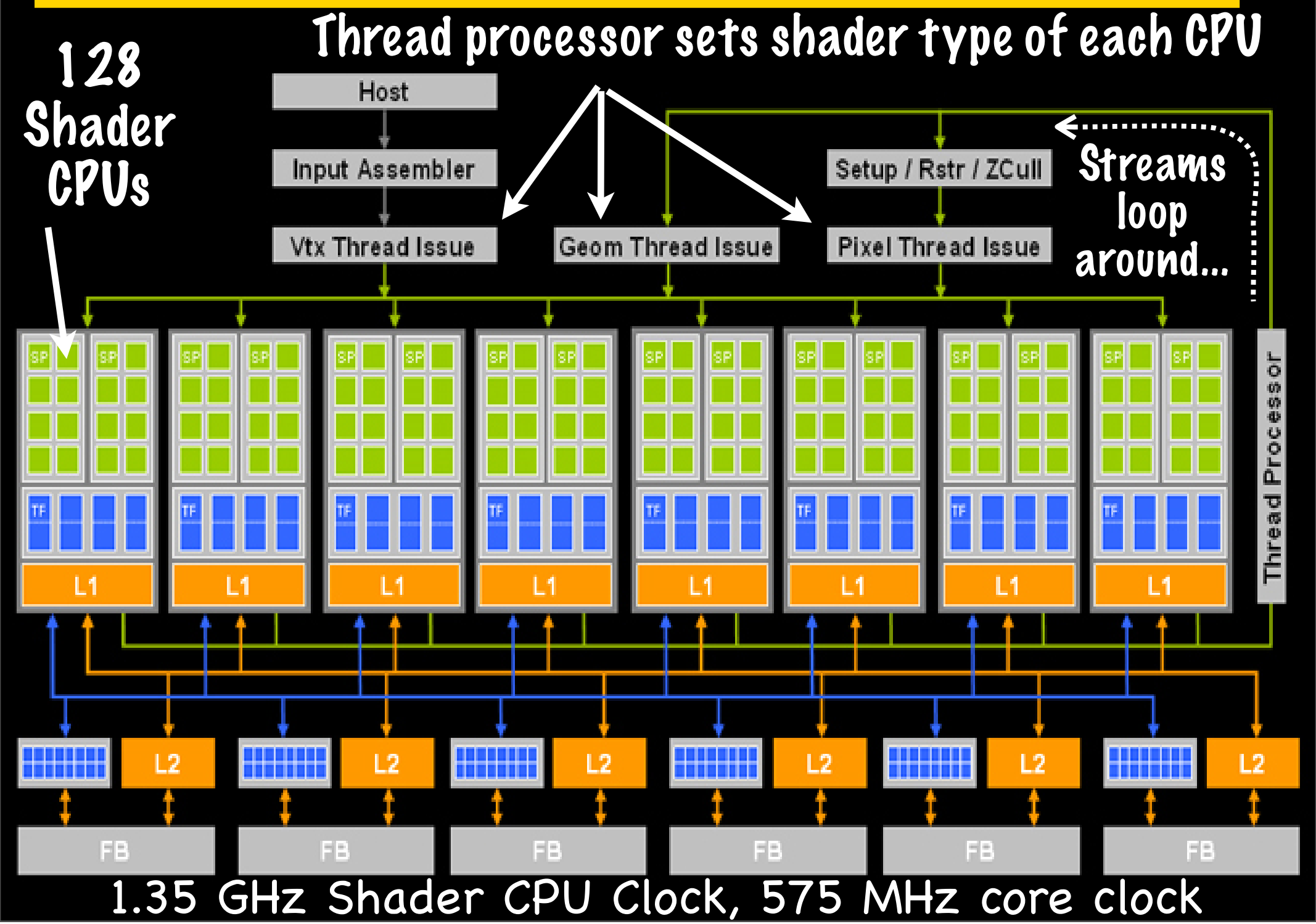
Why? Fractal images ...



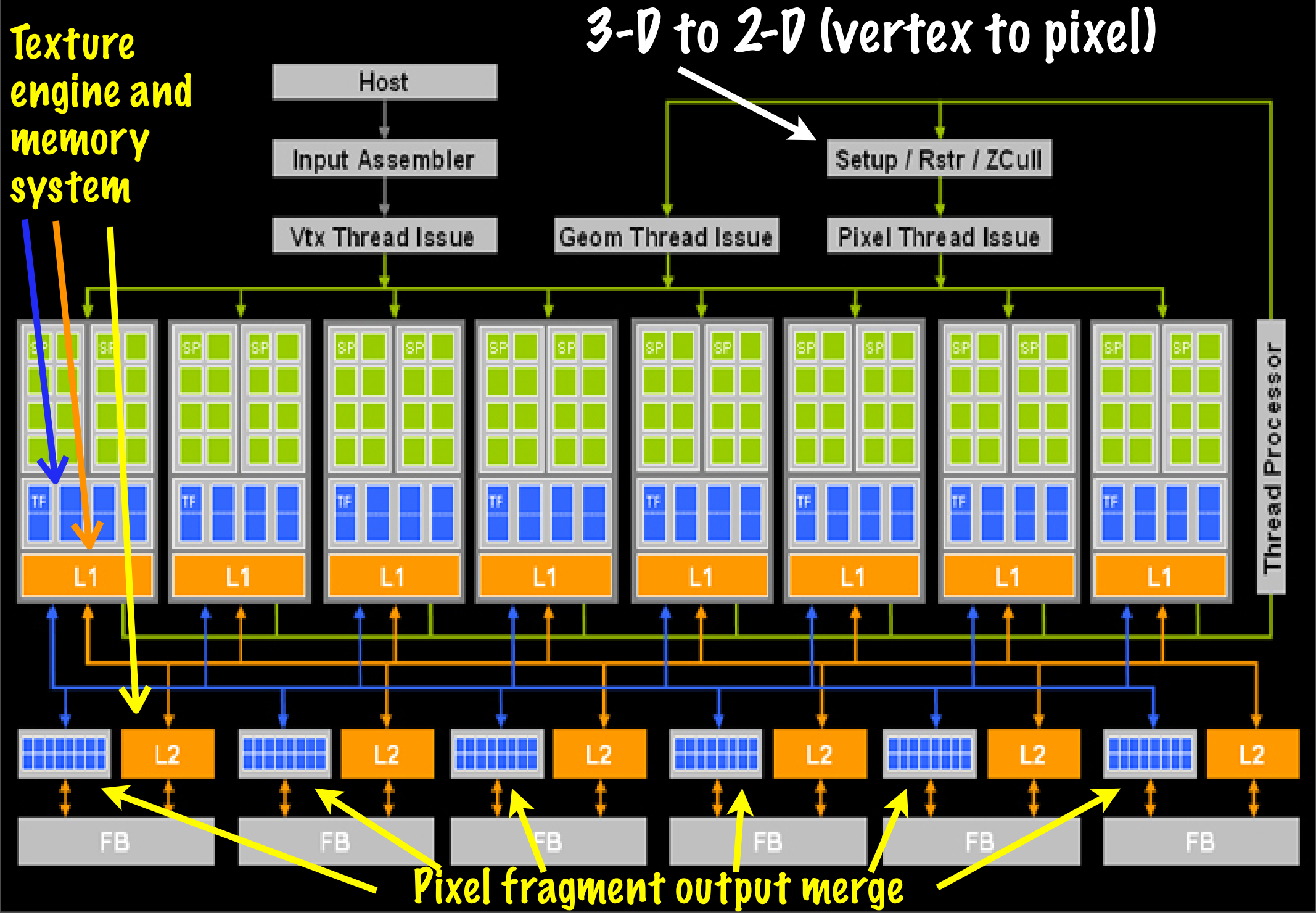
Why? Position-Based Fluids



NVidia 8800: Unified GPU, announced Fall 2006



Graphics-centric functionality …

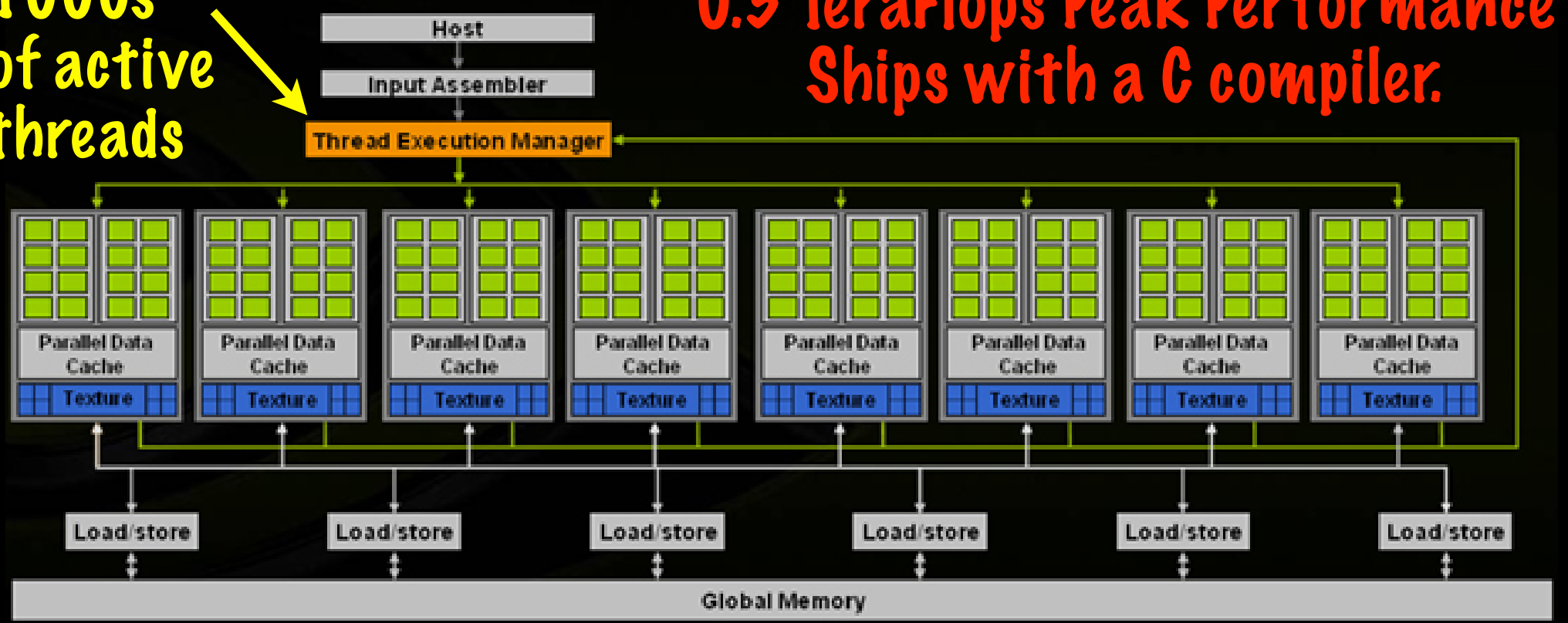


Can be reconfigured with graphics logic hidden …

128 scalar 1.35 GHz processors: Integer ALU, dual-issue single-precision IEEE floats.

1000s of active threads

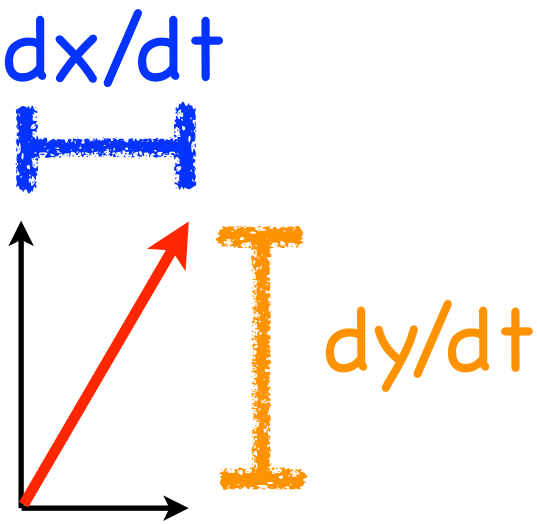
0.3 TeraFlops Peak Performance
Ships with a C compiler.

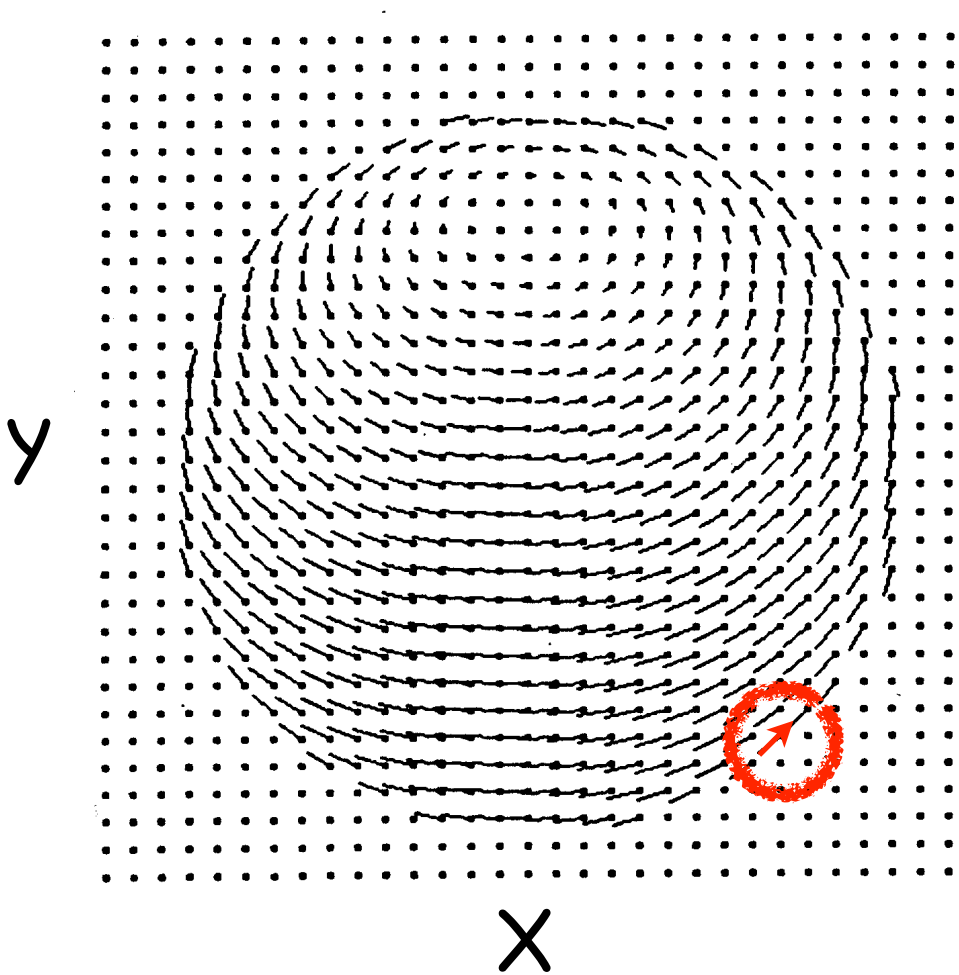


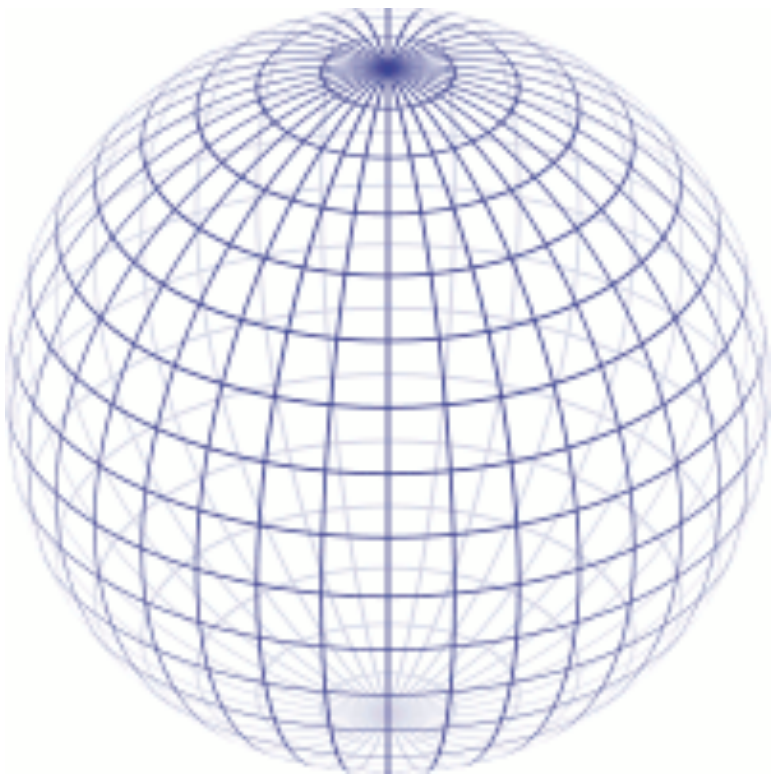
Texture system set up to look like a conventional memory system (768MB GDDR3, 86 GB/s)

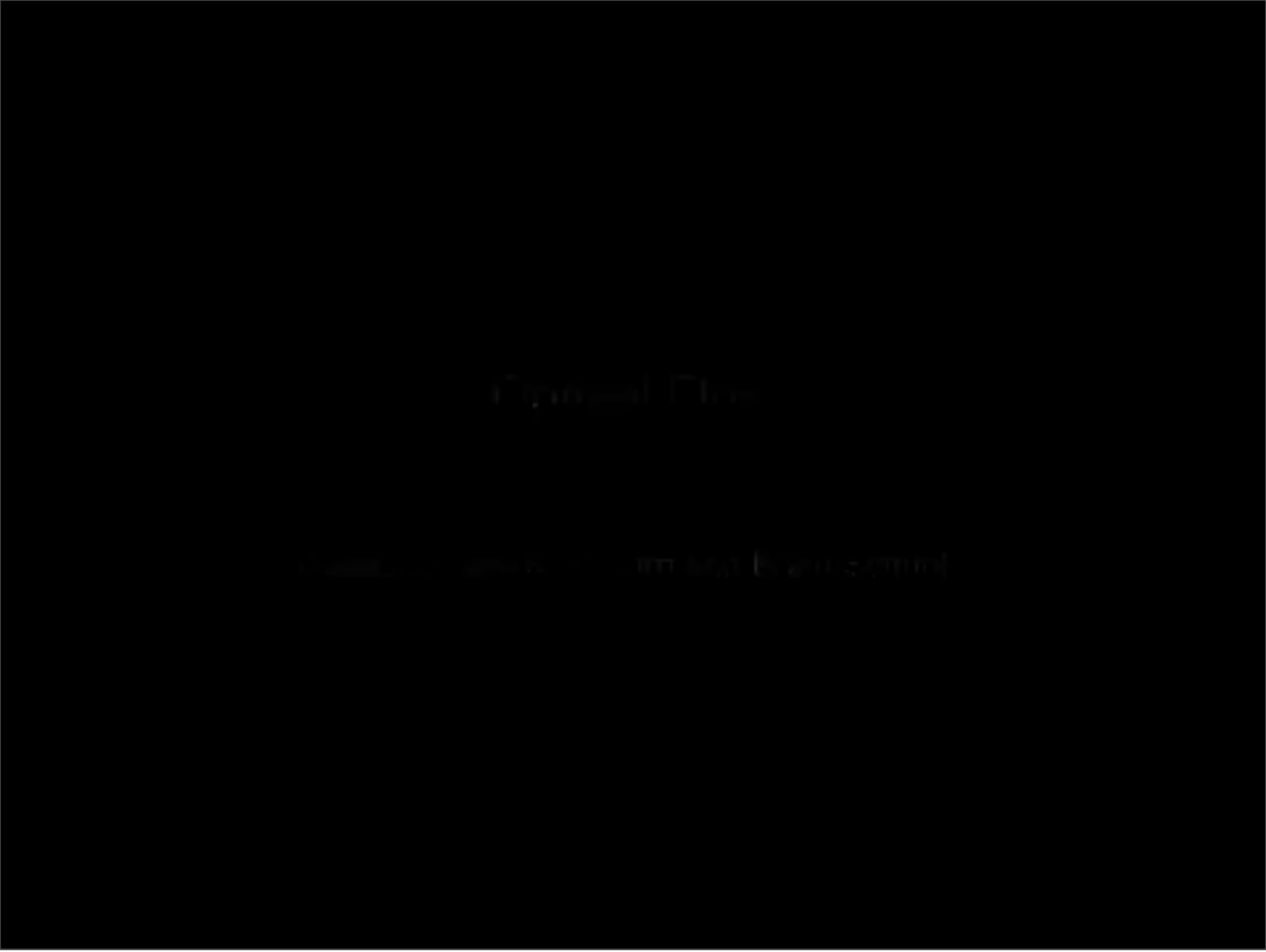
Optic Flow (Computer Vision)

Notate a movie with arrows to show speed and direction.









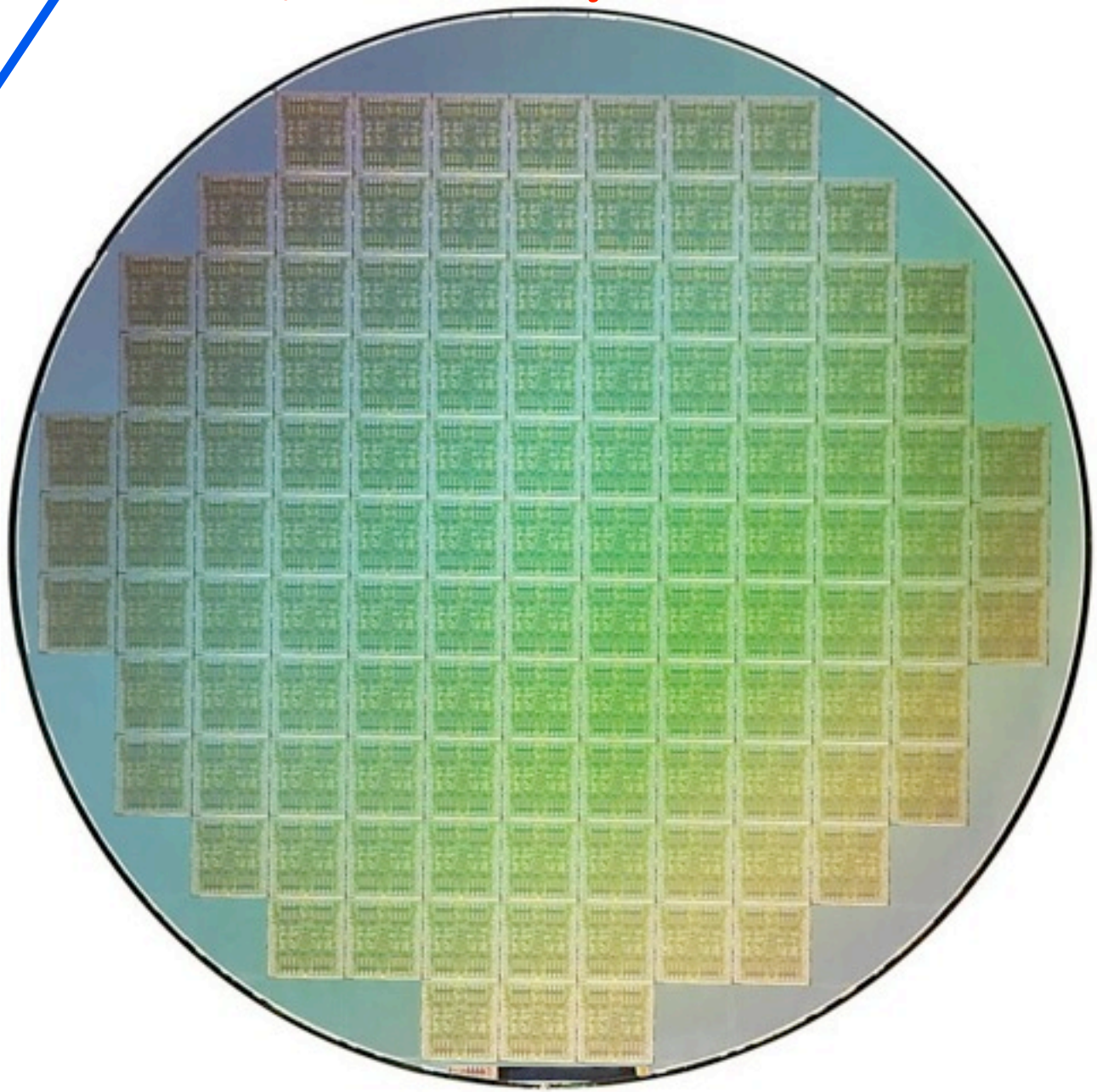
Chip Facts

90nm process

681M Transistors

80 die/wafer (pre-testing)

A big die. Many chips will not work (low yield). Low profits.



Design Facts

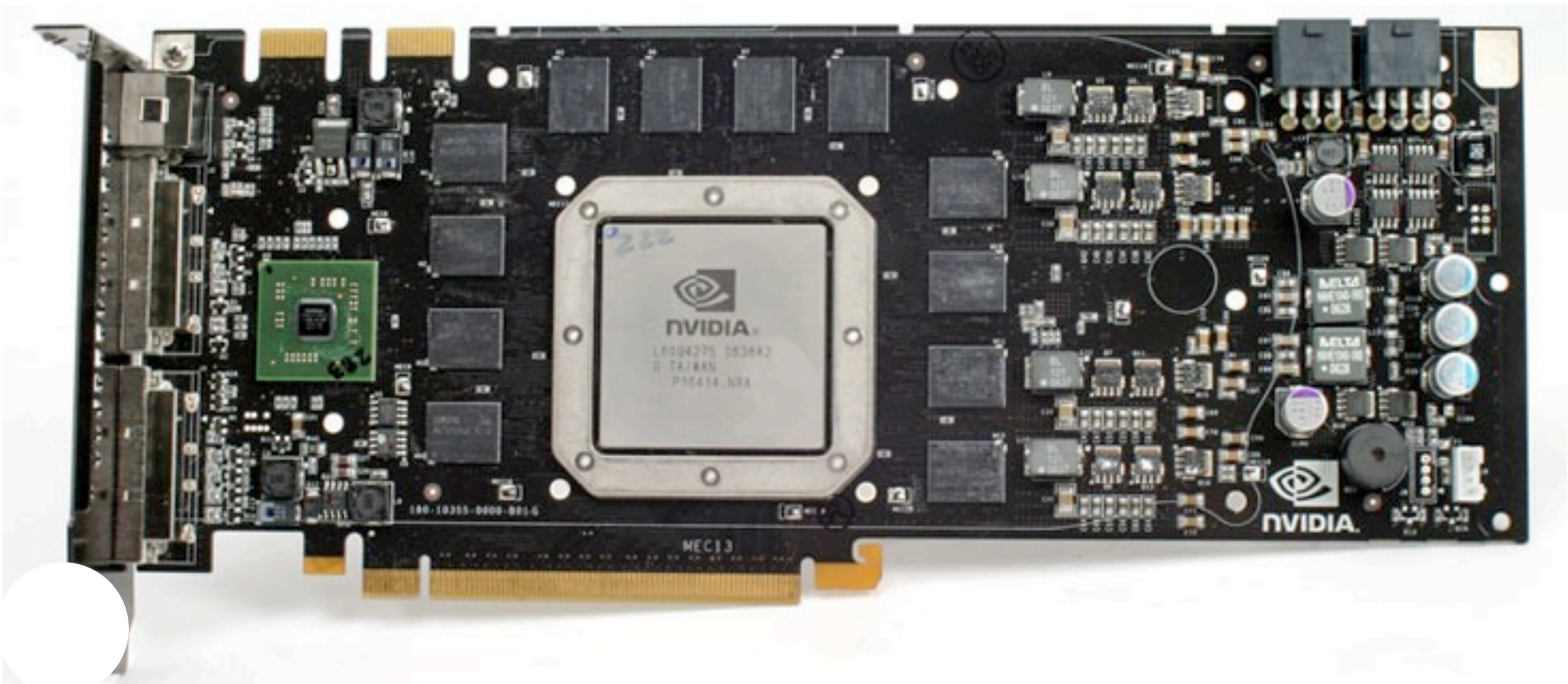
4 year design cycle

$400 Million design budget

600 person-years: 10 people at start, 300 at peak

GeForce 8800 GTX Card: $599 List Price

PCI-Express 16X Card - 2 Aux Power Plugs!



185 Watts Thermal Design Point (TDP) --
TDP is a "real-world" maximum power spec.

Some products are "loss-leaders"

Breakthrough product creates "free" publicity you can't buy.


Nvidia's Powerful Chip Moves Closer to 'Reality' - WSJ.com

file:///Users/lazzaro/Desktop/wsj-nvidia.webarchive

WSJ.com THE WALL STREET JOURNAL. ONLINE

Article Search

Advanced SearchAdvanced Search

Free Dow Jones Sites | Free Dow Jones Sites | As of Thursday, November 9, 2006 | Online Journ

Home
News
Technology
Markets
Personal Journal
Weekend & Leisure
Opinion
TODAY'S NEWSPAPER
MY ONLINE JOURNAL

Nvidia's Powerful Chip Moves Closer to 'Reality'

Graphics Product Captures Moving Shoulder Blades, Adrianne's Authentic Pout

By DON CLARK

November 9, 2006; Page B3

(1) Hope: when chip "shrinks" to 65nm fab process, die will be smaller, yields will improve, profits will rise.

(2) Simpler versions of the design will be made to create an entire product family, some very profitable.
"We tape out a chip a month", NVidia CEO quote.

And it happened! 2008 nVidia products

GTX 280

Price similar to 8800, stream CPU count > 2X.

	GTX 280	GTX 260	9800 GX2	9800 GTX+	9800 GTX
Stream Processors	240	192	256	128	128
Texture Address / Filtering	80 / 80	64 / 64	128 / 128	64 / 64	64 / 64
ROPs	32	28	32	16	16
Core Clock	602MHz	576MHz	600MHz	738MHz	675MHz
Shader Clock	1296MHz	1242MHz	1500MHz	1836MHz	1690MHz
Memory Clock	1107MHz	999MHz	1000MHz	1100MHz	1100MHz
Memory Bus Width	512-bit	448-bit	256-bit x 2	256-bit	256-bit
Frame Buffer	1GB	896MB	1GB	512MB	512MB
Transistor Count	1.4B	1.4B	1.5B	754M	754M
Manufacturing Process	TSMC 65nm	TSMC 65nm	TSMC 65nm	TSMC 55nm	TSMC 65nm
Price Point	$650	$400	$500	$229	$199

9800 GTX

Specs similar to 8800, card sells for $199.

And again in 2012! GTX 680 -- "Kepler"

GTX 680

3X more effective CPUs as GTX 280, lower price point.

6X more CPUs as 8800, (from 2006).

	GTX 680	GTX 580	GTX 560 Ti
Stream Processors	1536	512	384
Texture Units	128	64	64
ROPs	32	48	32
Core Clock	1006MHz	772MHz	822MHz
Shader Clock	N/A	1544MHz	1644MHz
Boost Clock	1058MHz	N/A	N/A
Memory Clock	6.008GHz GDDR5	4.008GHz GDDR5	4.008GHz GDDR5
Memory Bus Width	256-bit	384-bit	256-bit
Frame Buffer	2GB	1.5GB	1GB
FP64	1/24 FP32	1/8 FP32	1/12 FP32
TDP	195W	244W	170W
Transistor Count	3.5B	3B	1.95B
Manufacturing Process	TSMC 28nm	TSMC 40nm	TSMC 40nm
Launch Price	$499	$499	$249

GTX 560 Ti

Specs better than GTX 280, sells for $249

GTX 680

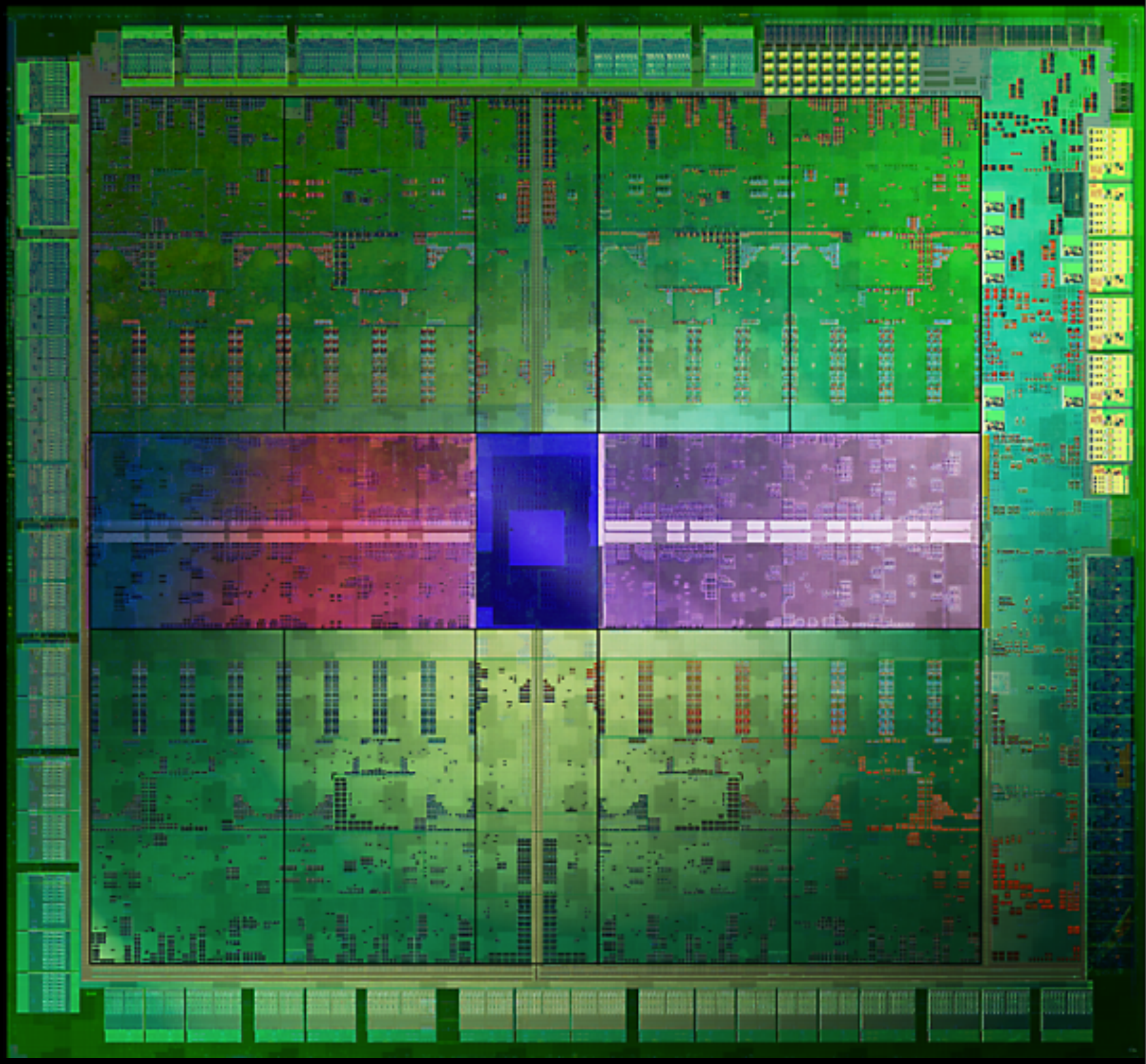
28nm process

3.5 billion transistors

1 GHz core clock

6GHz GDDR5

3 years, 1000 engineers



GTX 680

4X as many shader CPUs, running at 2/3 the clock (vs GTX 560).

Polymorph engine does polygon tessellation. PCIe bus no longer limits triangle count.

2013 -- Waiting for 20nm to arrive ...

GTX 780 Ti

1.6X more effective CPUs as GTX 770, 1.7x higher price.

Why? Still at 28 nm, so die size is larger.

	GTX 780 Ti	GTX 780	GTX 770
Stream Processors	2880	2304	1536
Texture Units	240	192	128
ROPs	48	48	32
Core Clock	875MHz	863MHz	1046MHz
Boost Clock	928Mhz	900Mhz	1085MHz
Memory Clock	7GHz GDDR5	6GHz GDDR5	7GHz GDDR5
Memory Bus Width	384-bit	384-bit	256-bit
VRAM	3GB	3GB	2GB
FP64	1/24 FP32	1/24 FP32	1/24 FP32
TDP	250W	250W	230W
Transistor Count	7.1B	7.1B	3.5B
Manufacturing Process	TSMC 28nm	TSMC 28nm	TSMC 28nm
Launch Date	11/07/13	05/23/13	05/30/13
Launch Price	$699	$649	$399

GTX 770

Specs close to GTX 680, $100 cheaper.

History and Graphics Processors

- **Create standard model from common practice:** Wire-frame geometry, triangle rasterization, pixel shading.
- **Put model in hardware:** Block diagram of chip matches computer graphics math.
- **Evolve to be programmable:** At some point, it becomes hard to see the math in the block diagram.

"Wheel of reincarnation" -- Hardwired graphics hardware evolves to look like general-purpose CPU. Ivan Sutherland co-wrote a paper on this topic in 1968!


SONY
PS4
© Mon Dean | www.mon-dean.com

Sony PS 4

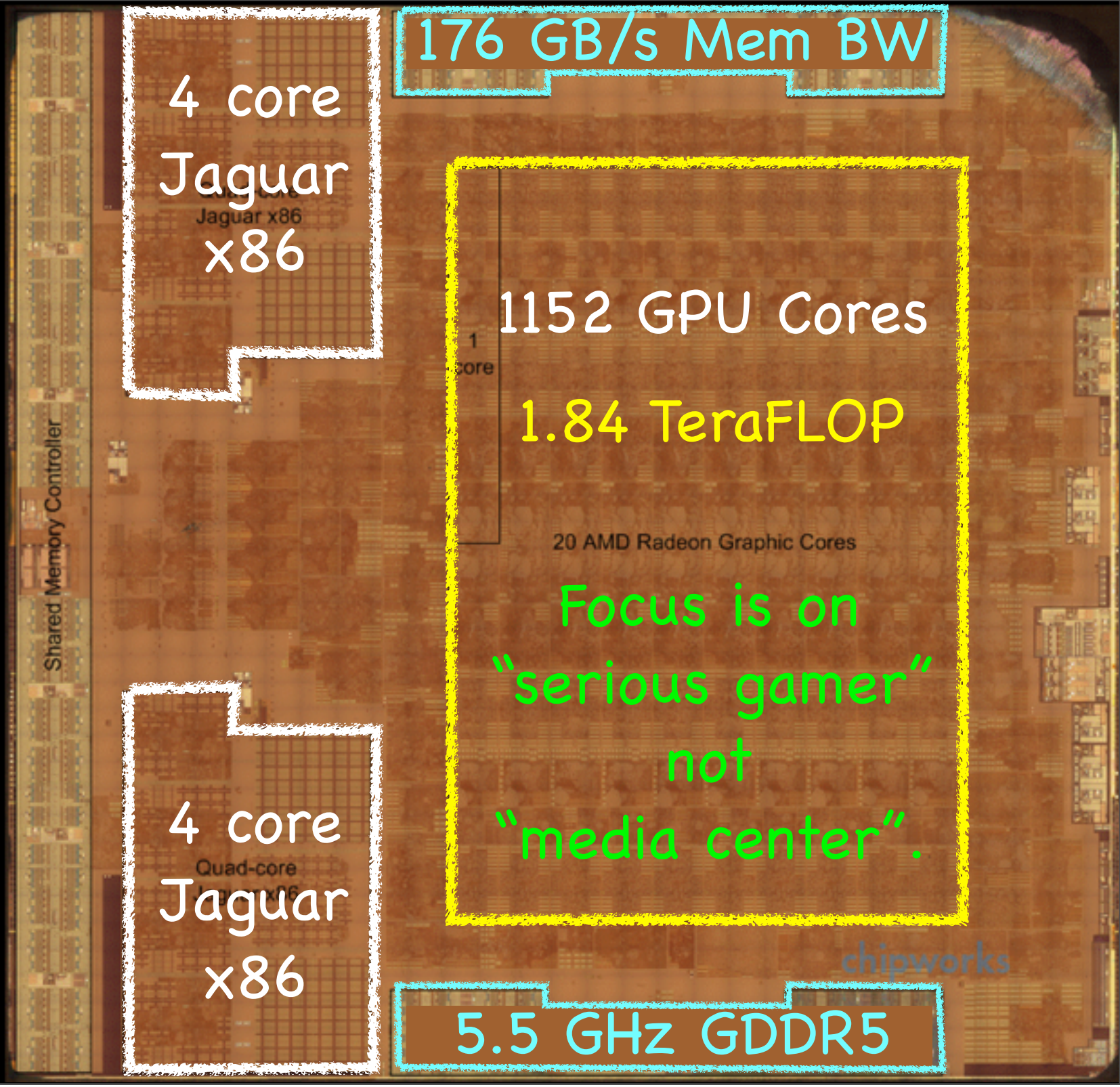
348 mm^2
28 nm HP

8GB
DRAM

140 W
system
power

$399 US

Camera
$59 extra




Steve Jobs
1955-2011

Apple A7

102 mm^2 die

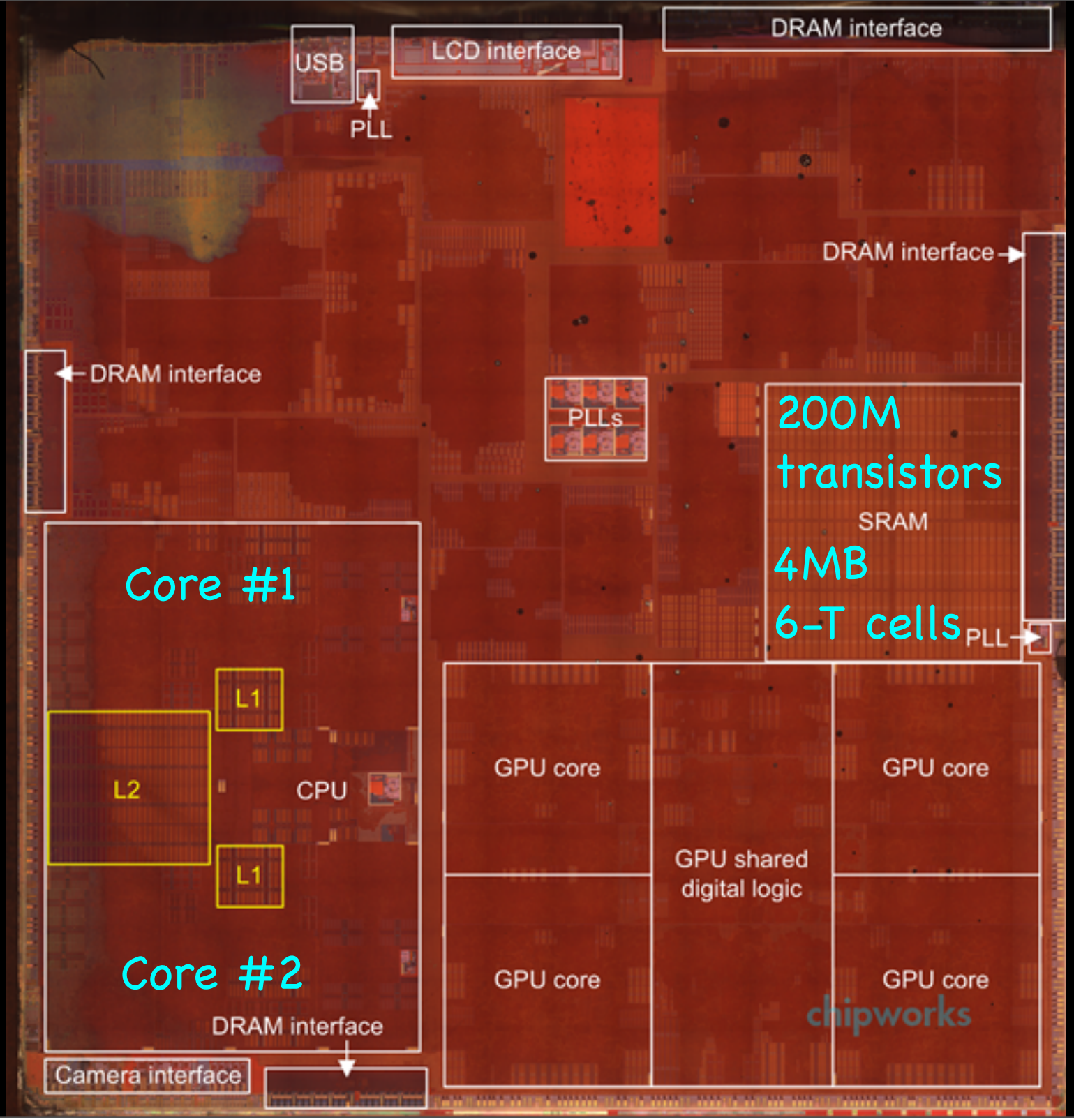
28 nm CMOS

1.1B transistors

64-bit ARM 1.4 GHz

GPU fills 22% of die

GPU: 2.7% of GTX 780 Ti (in GFLOPs).



On Tuesday

How did we get here?

Or, maybe another architecture topic ...

	T 4/22	A History of Intel x86 Microarchitecture		Intel Software Manual

Have a good weekend !