

The Elegance of Brute Force

Kurt Akeley

Graphics Architect

NVIDIA Corporation

GDC Europe, 26 August 2003

Outline

- **Performance Trends**
- **Brute Force**
- **Human Interface**



NVIDIA.

Performance

NVIDIA Performance History (AA 32-bit)

Season	Product	Mtri/sec	Yr rate	Mfrag/sec	Yr rate
2H97	Riva 128	3	-	20	-
1H98	Riva ZX	3	0.0	31	2.4
2H98	Riva TNT	6	4.0	50	2.6
1H99	Riva TNT2	9	2.3	75	2.3
2H99	GeForce256	15	2.8	120	2.6
1H00	GeForce2 GTS	25	2.8	200	2.8
2H00	GeForce2 Ultra	31	1.5	250	1.6
1H01	GeForce3	30	- 0.9	800	10.2
1H02	GeForce4 TI	60	2.0	1200	1.5
1H03	GeForce FX	200	3.3	2000	1.7

5.5 yrs

2.1

2.3

NVIDIA Performance History (No AA)

Season	Product	Mtri/sec	Yr rate	Mfrag/sec	Yr rate
2H97	Riva 128	5	-	100	-
1H98	Riva ZX	5	1.0	100	1.0
2H98	Riva TNT	5	1.0	180	3.2
1H99	Riva TNT2	8	1.0	333	3.4
2H99	GeForce	15	3.5	480	2.1
1H00	GeForce2 GTS	25	2.8	666	1.9
2H00	GeForce2 Ultra	31	1.5	1000	2.3
1H01	GeForce3	40	1.7	3200	10.2
1H02	GeForce4	65	1.6	4800	1.5

4.5 yrs

1.8

2.4



SGI Performance History (Depth Buffered)

Year	Product	Mtri/sec	Yr rate	Mfrag/sec	Yr rate
1984	Iris 2000	.0008	-	0.1	-
1988	GTX	.135	3.6	40	4.5
1992	RealityEngine	2.0	2.0	380	1.8
1996	InfiniteReality	12	1.6	1000	1.3

12 yrs

2.2

2.2



NVIDIA.

SGI Historical Performance (Flat Color)

Year	Product	Mtri/sec	Yr rate	Mfrag/sec	Yr rate
1984	Iris 2000	.010	-	46	-
1988	GTX	.135	1.9	80	1.2
1992	RealityEngine	2.0	2.0	380	1.5
1996	InfiniteReality	12	1.6	1000	1.3

12 yrs

1.8

1.3



NVIDIA.

Compound Performance Growth Rates

	Measured	Period	CAGR Tri / sec	CAGR Frag / sec
SGI	Flat Color	84 – 96	1.8	1.3
NVIDIA	No AA	97 – 02	1.8	2.4
SGI	Depth Buf	84 – 96	2.2	2.2
NVIDIA	AA 32-bit	97 – 03	2.1	2.3

Significantly above Moore's Law

CAGR 2.0 → 1000x per decade



Semiconductor Scaling Rates

From: *Digital Systems Engineering*, Dally and Poulton

Parameter	2001 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

Communication is the Key to Performance

- **Move data faster (optimize speed)**
 - Point-to-point wiring
 - Advanced protocols (e.g. clock in data)
 - Wide interfaces (256-bit GPUs)
- **Move data less (optimize locality)**
 - Algorithm
 - Architecture (e.g. pipeline GPU)
 - Cache data

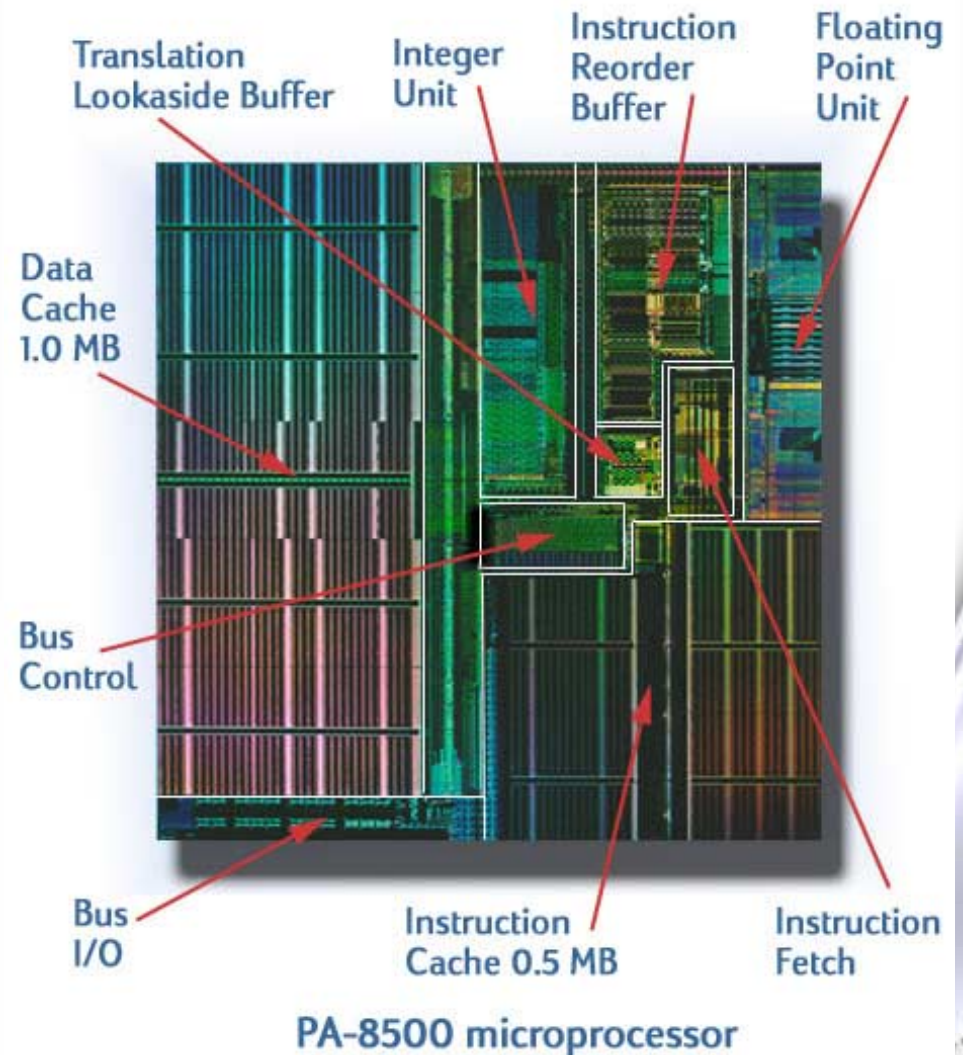


NVIDIA.

Microprocessors Are All Cache!

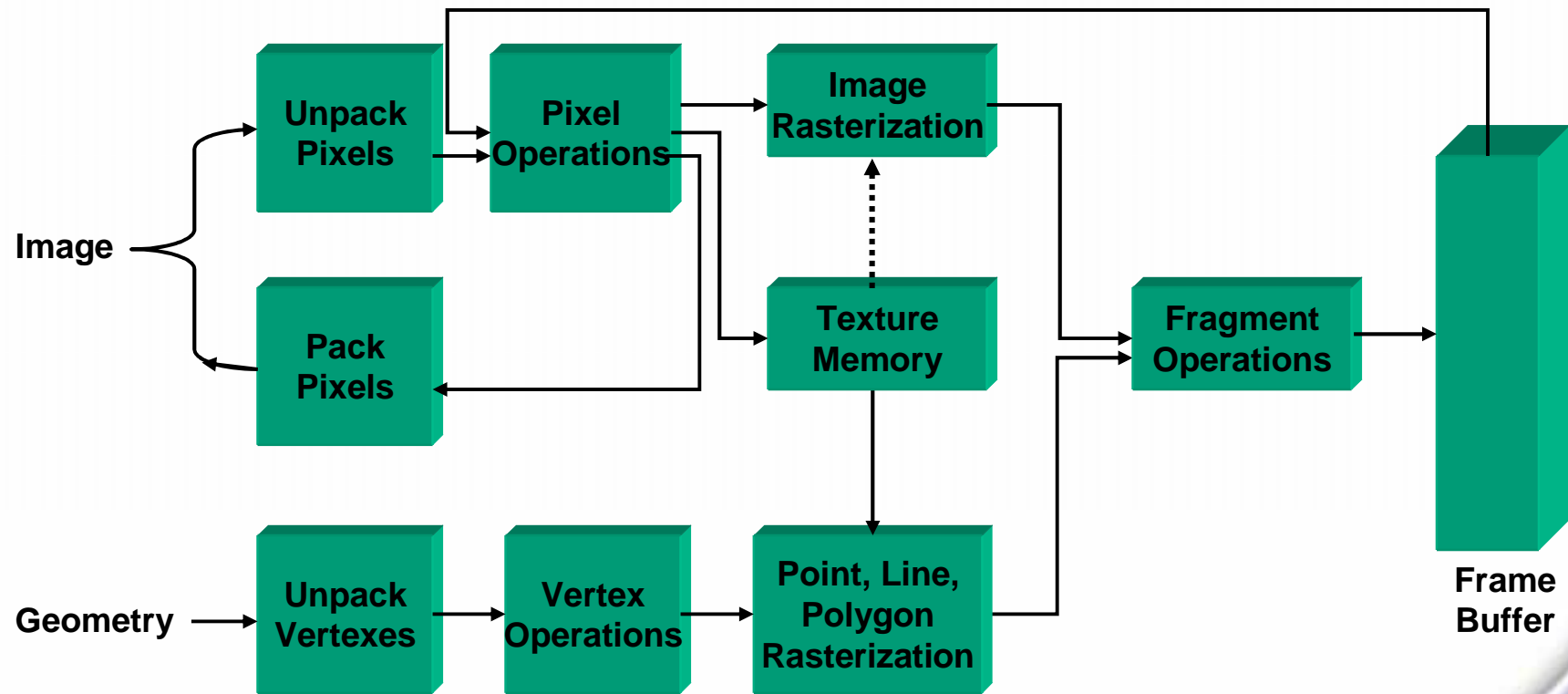
Locality optimized using cache memory

	CAGR	Growth in Decade
CPU →	1.5	58
	1.75	270
	2.0	1024
GPU →	2.25	3325
	2.5	9537

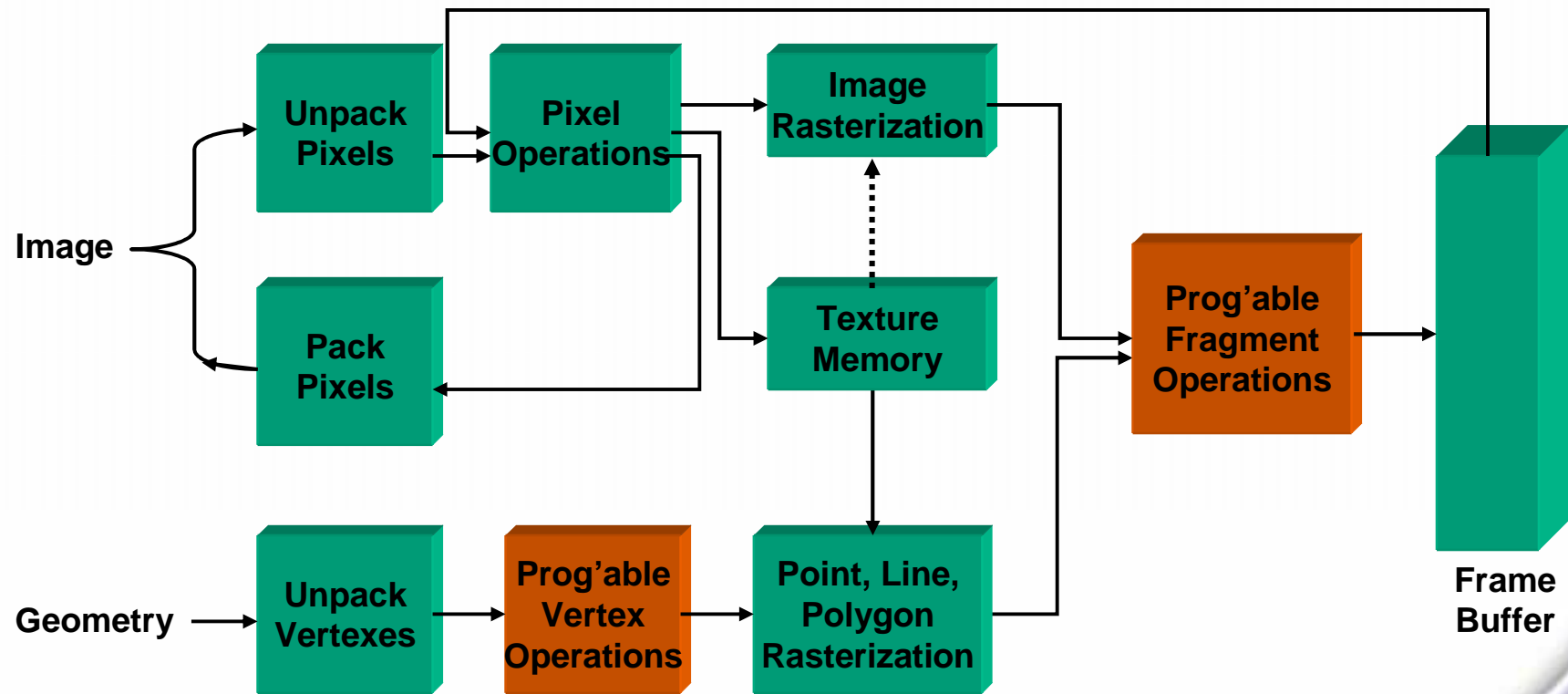


Brute Force

OpenGL 1992



OpenGL 2003



Graphics Pipeline

- **Locality optimized by algorithm / architecture**
 - Operate on individual vertexes
 - Operate on individual pixel fragments
 - Texture access is time-coherent
 - ...
- **Push model**
 - Little or no feedback to traversal
 - Data expansion (decompression)
- **Deep pipeline allows latency hiding**
 - Especially for RAM access (e.g. texture)



NVIDIA.

Depth Buffer – Elegant Brute Force

- **Properties**

- **Precise – exact at sample location**

- **Robust**

- **Sufficient**

- **Linear**

- **Within frame**

- **From frame to frame**

- **Locality**

- **NOT hidden surface elimination**

- **Nothing is ever determined about a surface**

- **No data reduction (except occlusion queries)**



NVIDIA.

Bottom Line

- **Depth buffer**
 - **Strong locality, highly parallel**
 - **Great for GPUs**
 - **Poor choice for CPUs**

- **Analytic hidden surface algorithm**
 - **Poor locality, not easily parallelized**
 - **Best choice for CPUs**
 - **Poor choice for GPUs**



NVIDIA.

“Great Game Graphics ... Who Cares?”

- GDC Europe Talk Title, 2003



NVIDIA.

Human Interface

Latency

- For an out-the-window display
 - 100 to 150 milliseconds
- For a head-mounted display
 - **5 to 15 milliseconds!**
- Total response latency, sum of
 - Tracking/input delay, plus
 - Rendering delay, plus
 - Display delay
- A 72 Hz display refreshes every 14 ms



NVIDIA.

Latency Solution

- **Reduce system latency to 5-15 ms range**
- **Requires 2-4 ms frame time (250-500 Hz)**
 - **Assuming 3-frame latency**
- **Estimated cost: 5x**



NVIDIA.

Running Total

Cost	Feature	Notes
5x	Low Latency	Frame rate 250-500 Hz



NVIDIA.

Stereo Solution

- **Binocular disparity is a very strong visual cue**
- **Must render separately for each eye**
 - **Occlusion**
 - **View-dependent lighting (e.g. reflections, specularities)**
 - **Alternatives tend to be hacks**
- **Estimated cost: 2x**



NVIDIA.

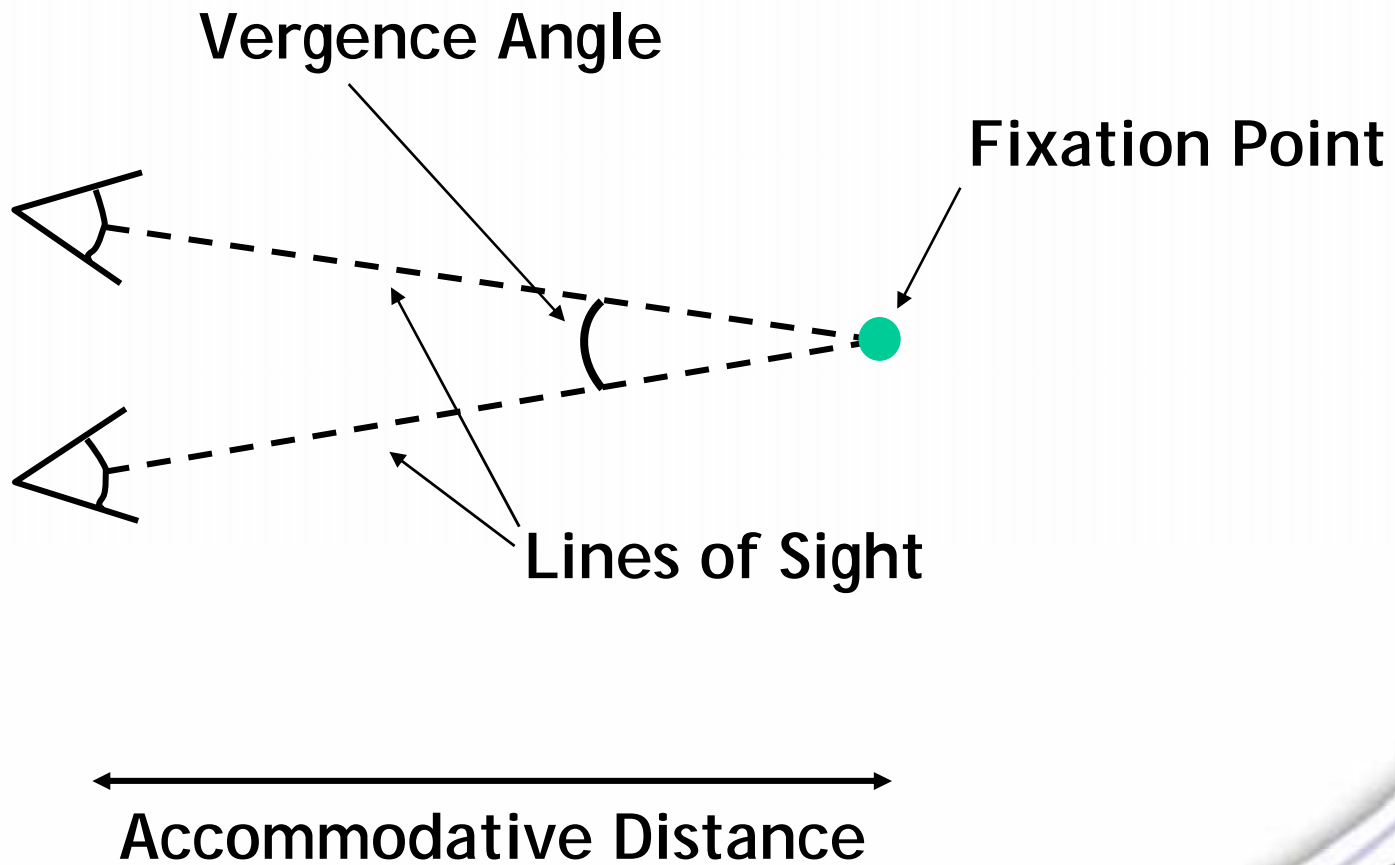
Running Total

Cost	Feature	Notes
5x	Low Latency	Frame rate 250-500 Hz
2x	Stereo	Two independent views

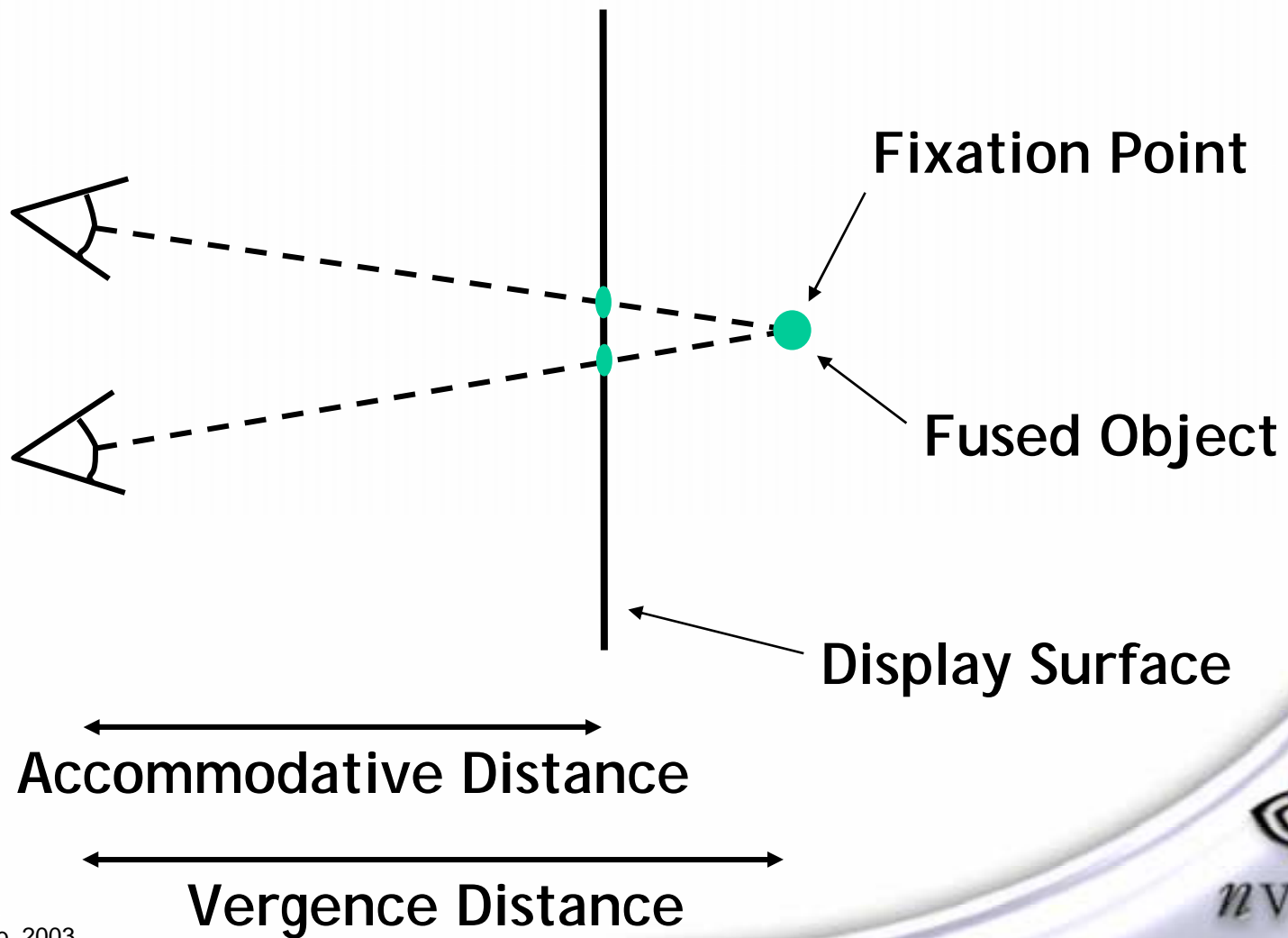


NVIDIA.

Vergence and Accommodation



Decoupling



Decoupling Causes ...

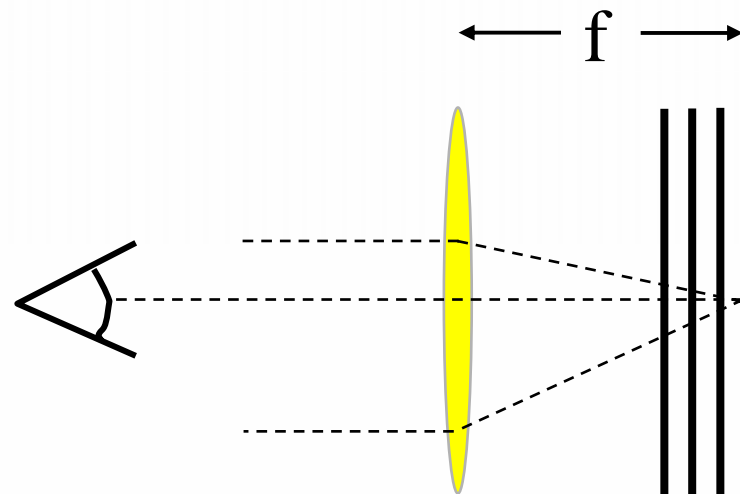
- **Incorrect estimations**
 - Distances
 - Angles?
- **Difficulty fusing stereo images**
 - Up to 2/3 of subjects unable to complete tasks
 - Random dot stereograms
- **Fatigue and discomfort**
 - Binocular Stress



NVIDIA.

Decoupling Solution

- **Volumetric display**
 - **Very low resolution in depth**
 - **Amounts to a 2.5D display**



- **Estimated cost: 3x**

Running Total

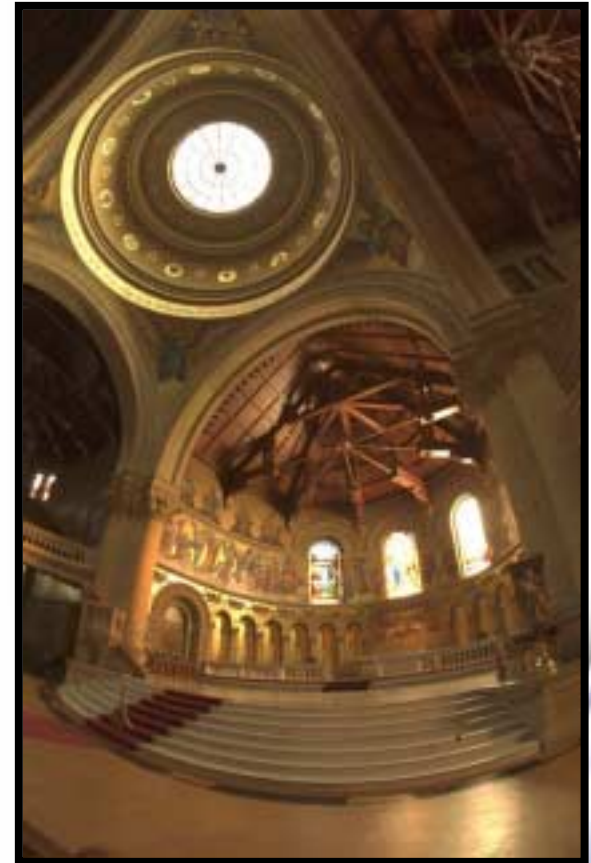
Cost	Feature	Notes
5x	Low Latency	Frame rate 250-500 Hz
2x	Stereo	Two independent views
3x	Correct Focus	Vergence and accommodation coupled



NVIDIA.

High Dynamic Range (HDR)

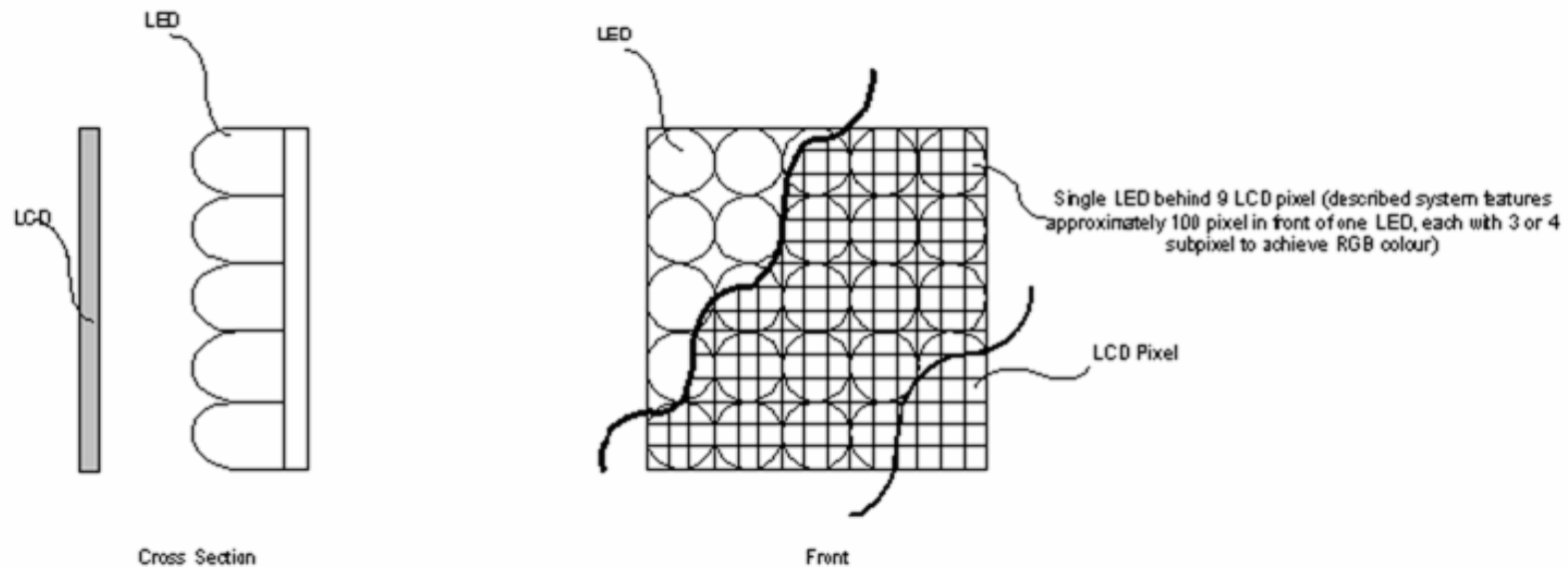
- **Human limitations**
 - 1,000,000:1 range of sensitivity
 - 100,000:1 contrast within scene
- **Current displays**
 - CRT 300:1 contrast ratio
 - LCD 500:1 contrast ratio
- **SIGGRAPH 2003 ET** →
 - Sunnybrook Technologies



Sunnybrook Technologies

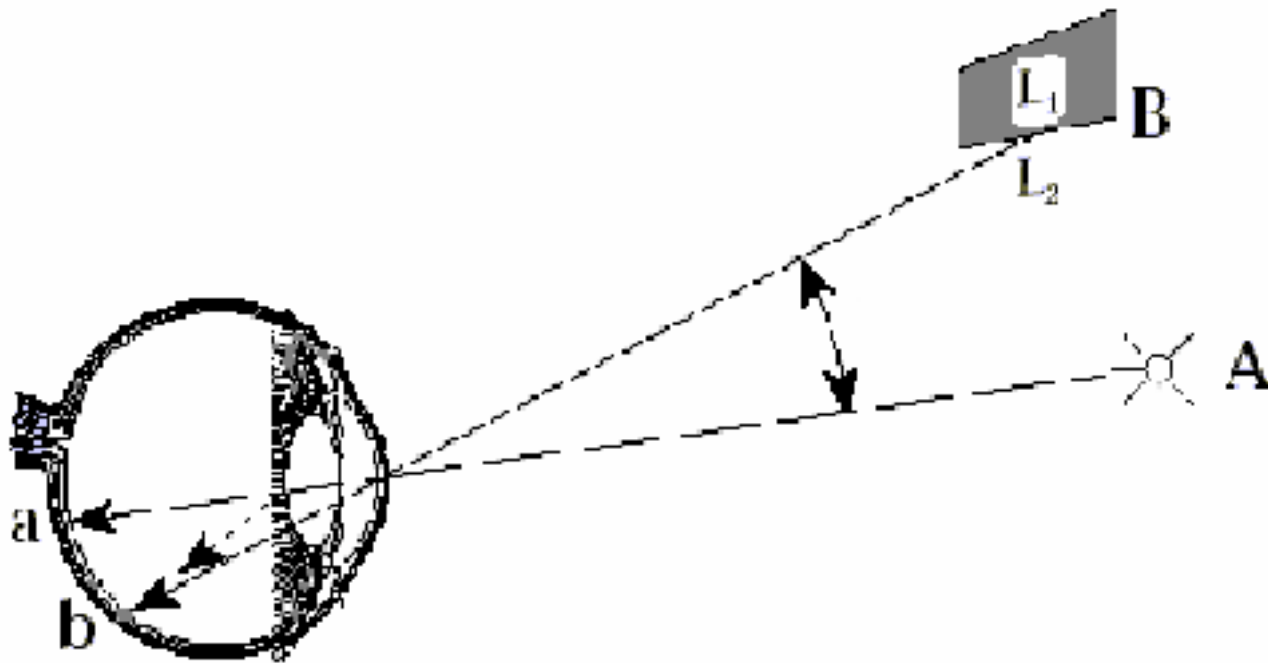
Dual-density display

- Conventional LCD panel in front (full-resolution)
- White LED array used as back-light (~1/50 resolution)



Sunnybrook Technologies

- Scattering masks low resolution LEDs



HDR Solution

- **Requires 16-bit framebuffer components**
 - **Rendering**
 - **Blending**
 - **Full-scene anti-aliasing**
- **Requires multi-resolution rendering**
 - **Full-resolution for LCD, corrected for back-lighting**
 - **Low-resolution for back-lighting**
- **Estimated cost: 2x**



NVIDIA.

Running Total

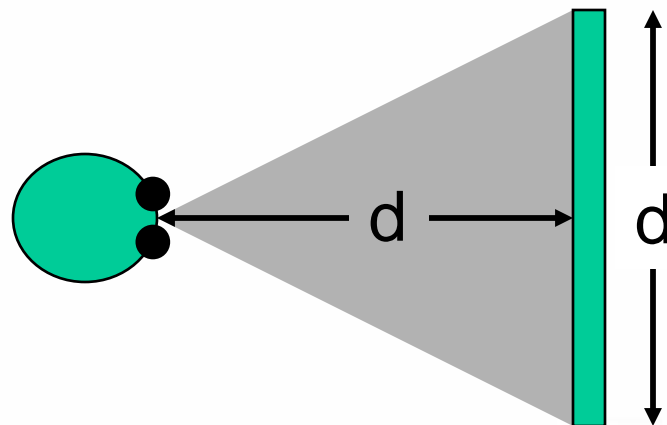
Cost	Feature	Notes
5x	Low Latency	Frame rate 250-500 Hz
2x	Stereo	Two independent views
3x	Correct Focus	Vergence and accommodation coupled
2x	HDR	Multi-resolution rendering



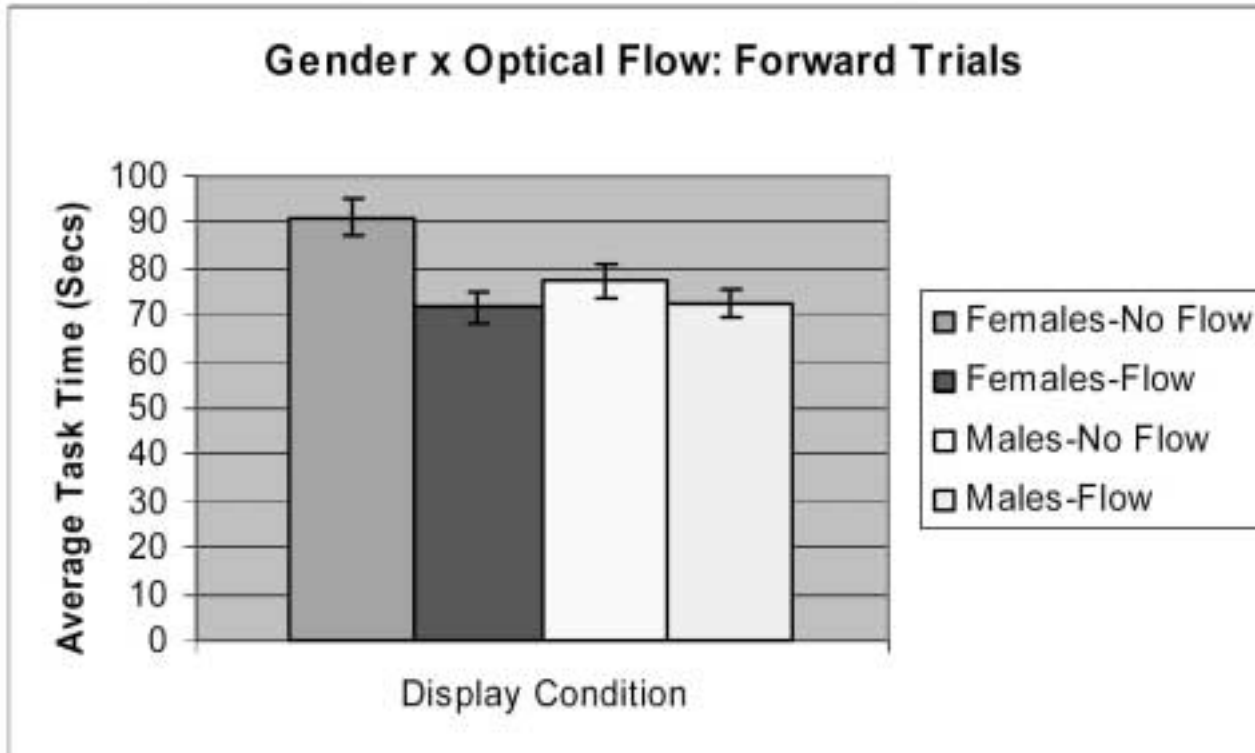
NVIDIA.

Field of View

- **Human field of view (FOV)**
 - **Monocular: 160 deg (wide) x 135 deg (high)**
 - **Binocular: 200 deg (wide)**
 - **Binocular overlap: 120 deg (wide)**
- **Typical screen FOV**
 - **55 deg (wide) x 41 deg (high)**



Optical Flow Matters



“Women Go With the (Optical) Flow”, Desney S. Tan, Mary Czerwinski, George Robertson.

<http://research.microsoft.com/users/marycz/chi2003flow.pdf>



FOV Solution

- **Double horizontal FOV to 110 degrees**
- **Double vertical FOV to 80 degrees**
- **Cleverness to distribute resolution ?**
 - e.g. cylindrical projection

- **Estimated cost: 4x**



NVIDIA.

Running Total

Cost	Feature	Notes
5x	Low Latency	Frame rate 250-500 Hz
2x	Stereo	Two independent views
3x	Correct Focus	Vergence and accommodation coupled
2x	HDR	Multi-resolution rendering
4x	Full FOV	110 deg (wide) x 80 deg (high)



NVIDIA.

Foveal Resolution

- **Foveal sampling density is $\frac{1}{2}$ arc minute**
 - 120 pixels / degree
 - Packing is roughly hexagonal
- **Typical monitor sampling is 2 arc minutes**
 - 1600 pixels at (dist = width)
- **IBM T221 (aka Big Bertha) LCD Display**
 - Resolution: 3840 (wide) x 2400 (high)
 - Dimensions: 19" (wide) x 12" (high)
- **Estimated cost: 15x**



NVIDIA.

Running Total

Cost	Feature	Notes
5x	Low Latency	Frame rate 250-500 Hz
2x	Stereo	Two independent views
3x	Correct Focus	Vergence and accommodation coupled
2x	HDR	Multi-resolution rendering
4x	Full FOV	110 deg (wide) x 80 deg (high)
15x	Foveal Resolution	1/2 arc minute resolution



NVIDIA.

Full-Scene Antialiasing

- **SAGE**
- **Render**
 - **16 sample / pixel**
- **Reconstruction**
 - **5x5 pixel filter**
 - **400 samples / pixel**
 - **~1000 FLOPs / pixel**
- **Estimated cost: 5x**



“The SAGE Graphics Architecture”, Michael Deering and David Naegle, Proceedings of SIGGRAPH 2002



Running Total

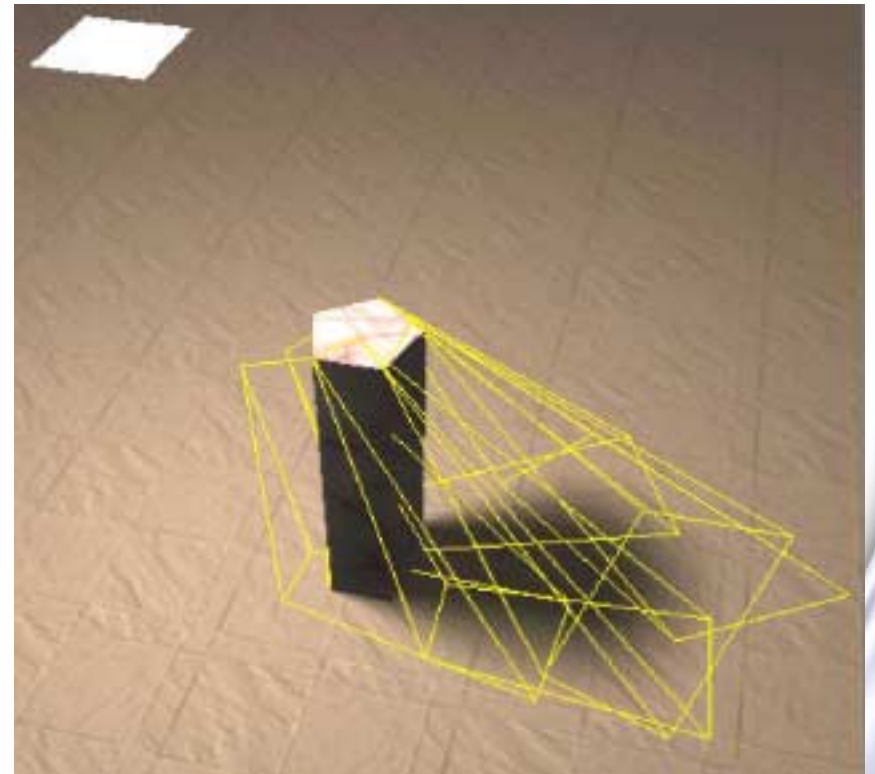
Cost	Feature	Notes
5x	Low Latency	Frame rate 250-500 Hz
2x	Stereo	Two independent views
3x	Correct Focus	Vergence and accommodation coupled
2x	HDR	Multi-resolution rendering
4x	Full FOV	110 deg (wide) x 80 deg (high)
15x	Foveal Resolution	1/2 arc minute resolution
5x	FSAA	16 samples / pixel, 5x5 pixel filter



NVIDIA.

Soft Shadows

- Look nice
- Help define spatial relationships
- Still expensive
- Estimated cost: 2x ?



“A Geometry-based Soft Shadow Volume Algorithm using Graphics Hardware”, Ulf Assarsson and Tomas Akenine-Möller, Proceedings of SIGGRAPH 2002



Running Total

Cost	Feature	Notes
5x	Low Latency	Frame rate 250-500 Hz
2x	Stereo	Two independent views
3x	Correct Focus	Vergence and accommodation coupled
2x	HDR	Multi-resolution rendering
4x	Full FOV	110 deg (wide) x 80 deg (high)
15x	Foveal Resolution	1/2 arc minute resolution
5x	FSAA	16 samples / pixel, 5x5 pixel filter
2x	Soft Shadows	Define spatial relationships



NVIDIA.

Let's Sum It All Up

Cost	Feature	Notes
5x	Low Latency	Frame rate 250-500 Hz
2x	Stereo	Two independent views
3x	Correct Focus	Vergence and accommodation coupled
2x	HDR	Multi-resolution rendering
4x	Full FOV	110 deg (wide) x 80 deg (high)
15x	Foveal Resolution	1/2 arc minute resolution
5x	FSAA	16 samples / pixel, 5x5 pixel filter
2x	Soft Shadows	Define spatial relationships

36,000x



This Will Keep Us Busy ...

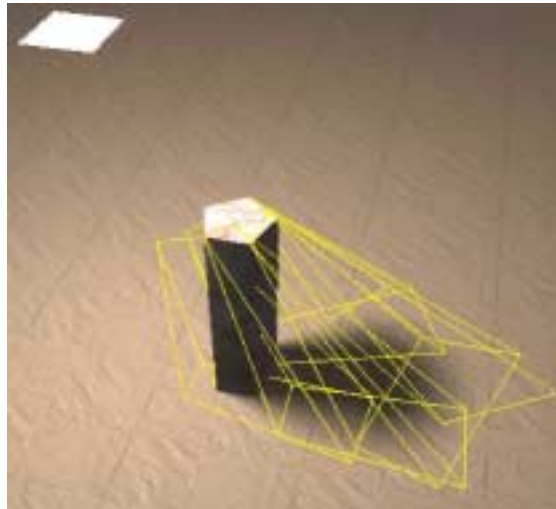
Multiple	2.2 CAGR	2.0 CAGR	1.8 CAGR
1000	9 years	10 years	12 years
5000	11 years	12 years	15 years
10000	12 years	13 years	16 years
50000	15 years	16 years	18 years

36,000x



It's Not Over Yet

- Lots of performance headroom
- Lots of performance need
 - Human interface
 - Better images too ...



NVIDIA.