What Next?

Kevin Walsh
CS 3410, Spring 2010
Computer Science
Cornell University

* slides thanks to Kavita Bala & many others

Final Project

Demo Sign-Up:

- Will be posted outside my office after lecture today.

10 time slots on Wednesday, May 19

24 time slots on Thursday, May 20

12 time slots on Friday, May 21

CMS submission due:

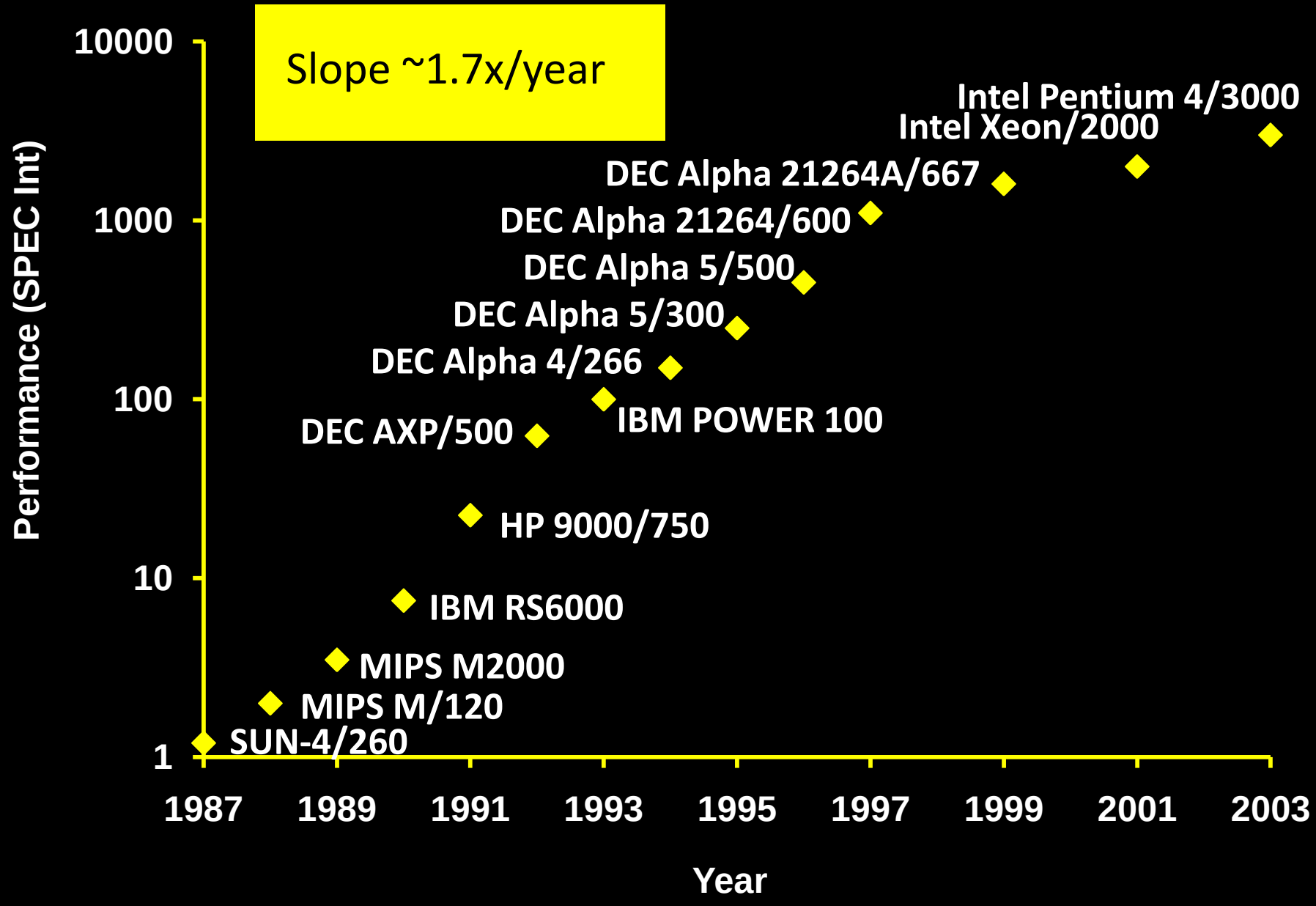
- Afternoon of Friday, May 21

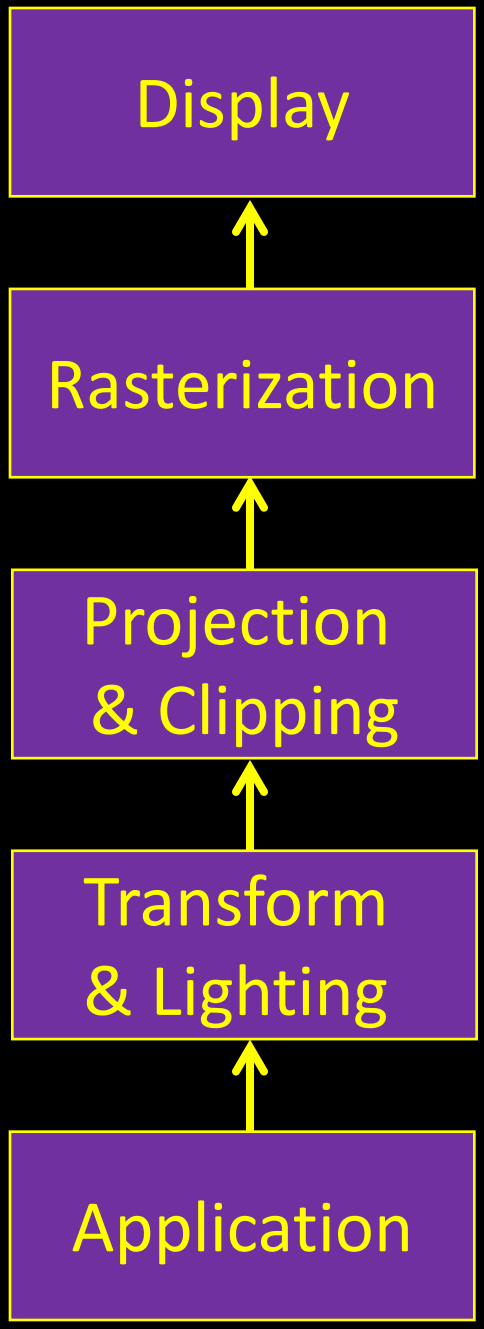
Moore's Law introduced in 1965

- Number of transistors that can be integrated on a single die would double every 18 to 24 months (i.e., grow exponentially with time).

Amazingly visionary

- 2300 transistors, 1 MHz clock (Intel 4004) - 1971
- 16 Million transistors (Ultra Sparc III)
- 42 Million transistors, 2 GHz clock (Intel Xeon) – 2001
- 55 Million transistors, 3 GHz, 130nm technology, 250mm2 die (Intel Pentium 4) – 2004
- 290+ Million transistors, 3 GHz (Intel Core 2 Duo) – 2007


Slope ~1.7x/year
Intel Pentium 4/3000
Intel Xeon/2000
DEC Alpha 21264A/667
DEC Alpha 21264/600
DEC Alpha 5/500
DEC Alpha 5/300
DEC Alpha 4/266
IBM POWER 100
DEC AXP/500
HP 9000/750
IBM RS6000
MIPS M2000
MIPS M/120
SUN-4/260
10000
1000
100
10
1
Performance (SPEC Int)
1987
1989
1991
1993
1995
1997
1999
2001
2003
Year



The dark ages (early-mid 1990's), when there were only frame buffers for normal PC's.

Some accelerators were no more than a simple chip that sped up linear interpolation along a single span, so increasing fill rate.

This is where pipelines start for PC commodity graphics, prior to Fall of 1999.

This part of the pipeline reaches the consumer level with the introduction of the NVIDIA GeForce256.

Hardware today is moving traditional application processing (surface generation, occlusion culling) into the graphics accelerator.

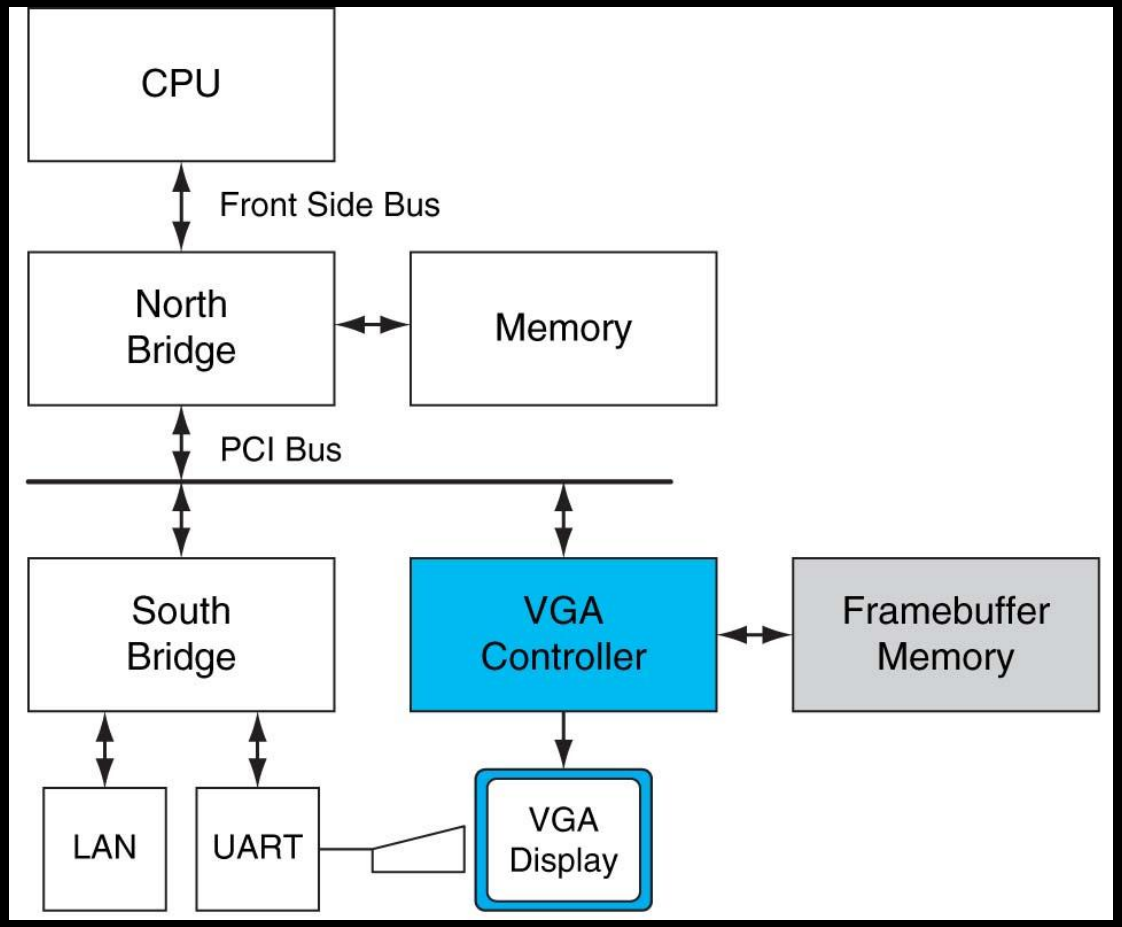
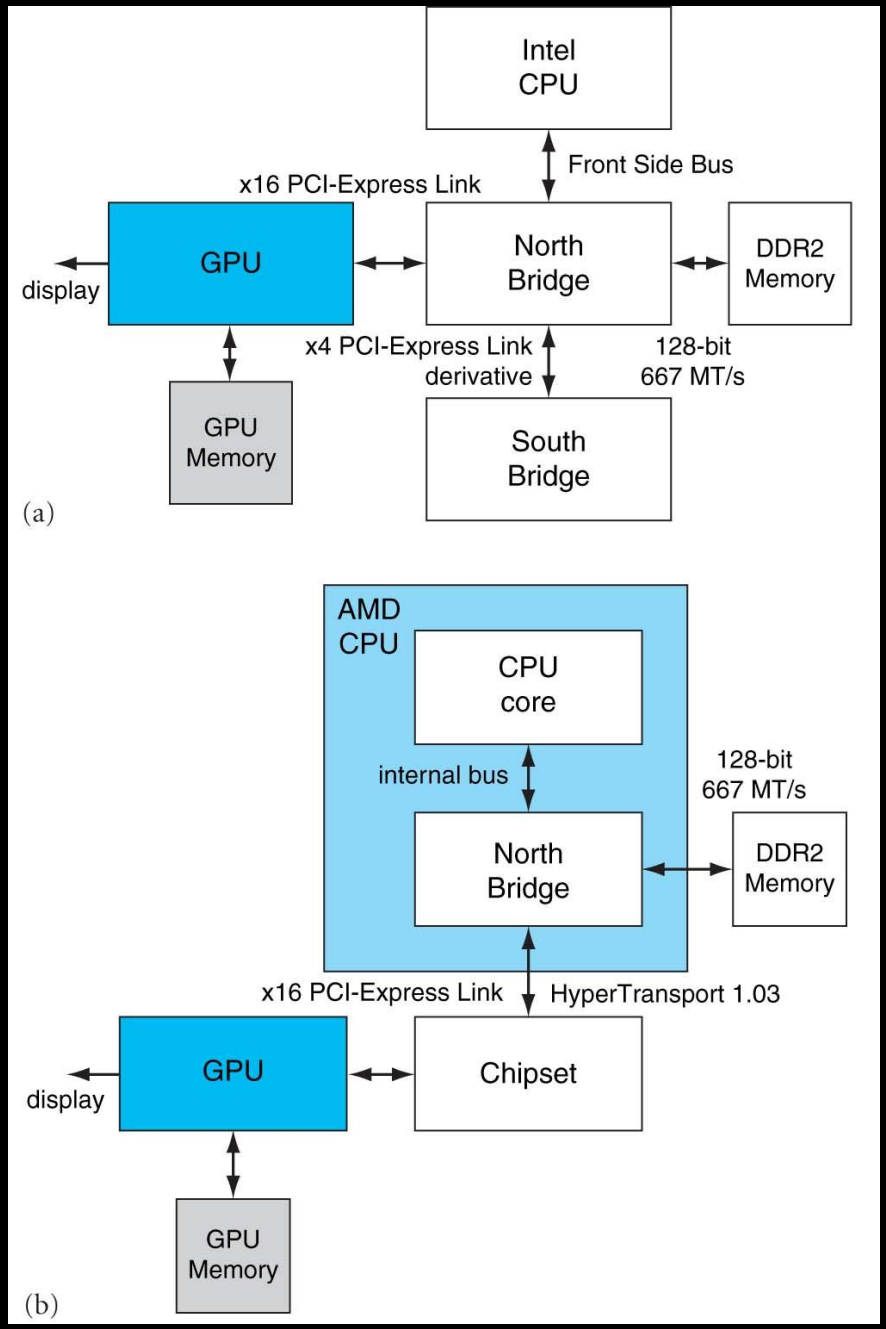
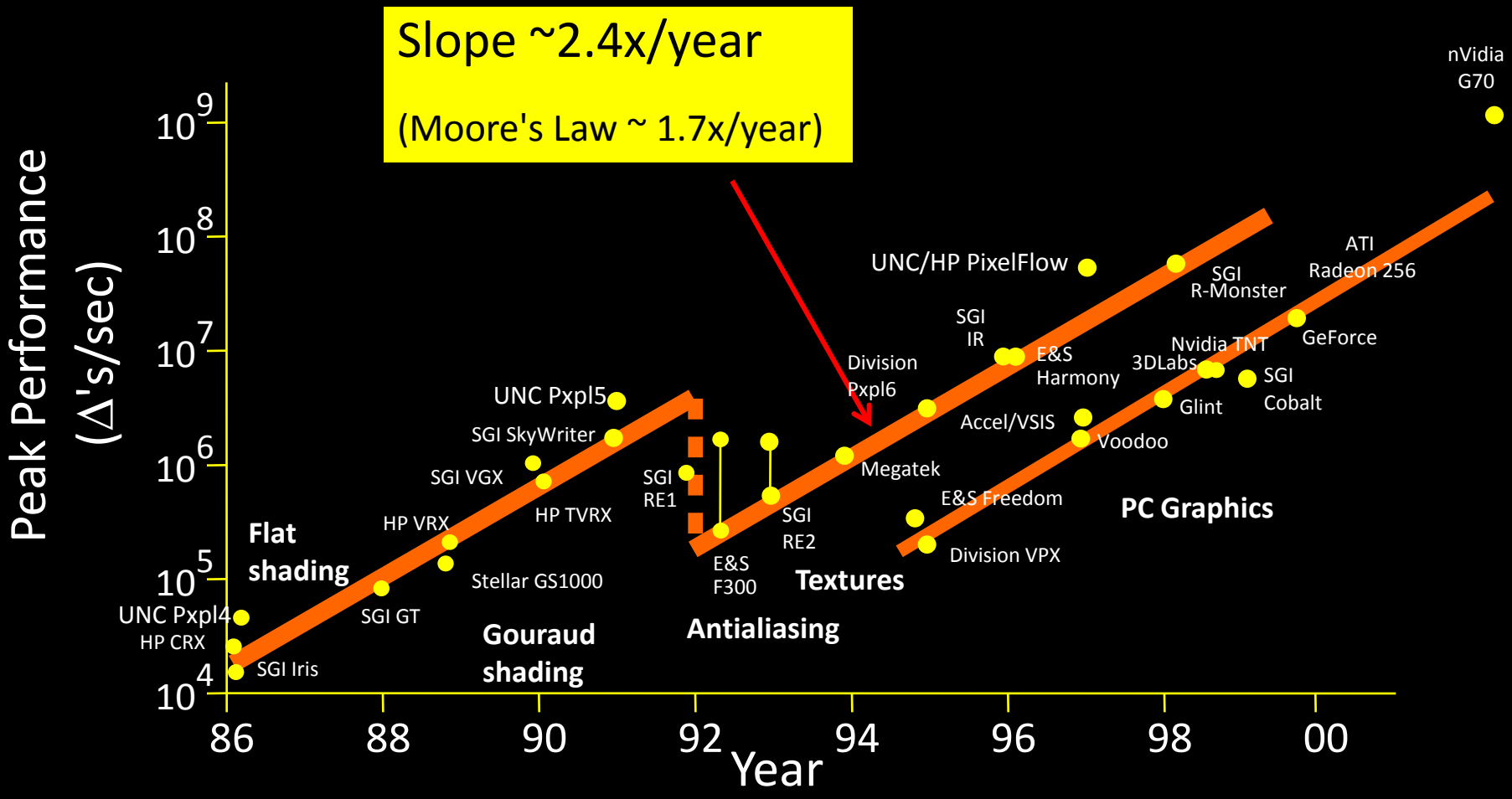


FIGURE A.2.1 Historical PC. VGA controller drives graphics display from framebuffer memory.


Intel
CPU
Front Side Bus
x16 PCI-Express Link
GPU
North
Bridge
DDR2
Memory
display
x4 PCI-Express Link
derivative
128-bit
667 MT/s
GPU
Memory
South
Bridge
(a)
AMD
CPU
CPU
core
internal bus
128-bit
667 MT/s
North
Bridge
DDR2
Memory
x16 PCI-Express Link
HyperTransport 1.03
GPU
Chipset
display
GPU
Memory
(b)

One-pixel polygons (~10M polygons @ 30Hz)



Graph courtesy of Professor John Poulton (from Eric Haines)


NVIDIA
TESLA

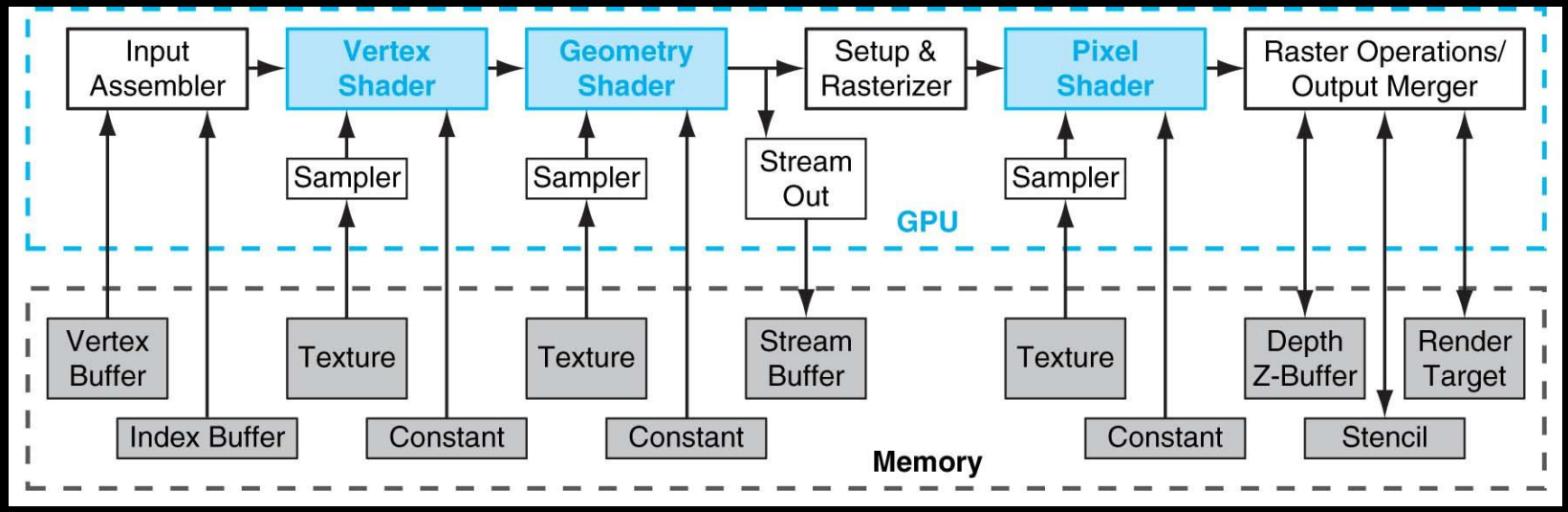


FIGURE A.3.1 Direct3D 10 graphics pipeline. Each logical pipeline stage maps to GPU hardware or to a GPU processor. Programmable shader stages are blue, fixed-function blocks are white, and memory objects are grey. Each stage processes a vertex, geometric primitive, or pixel in a streaming dataflow fashion.

Pipelined and parallel

Very, very parallel: 128 to 1000 cores

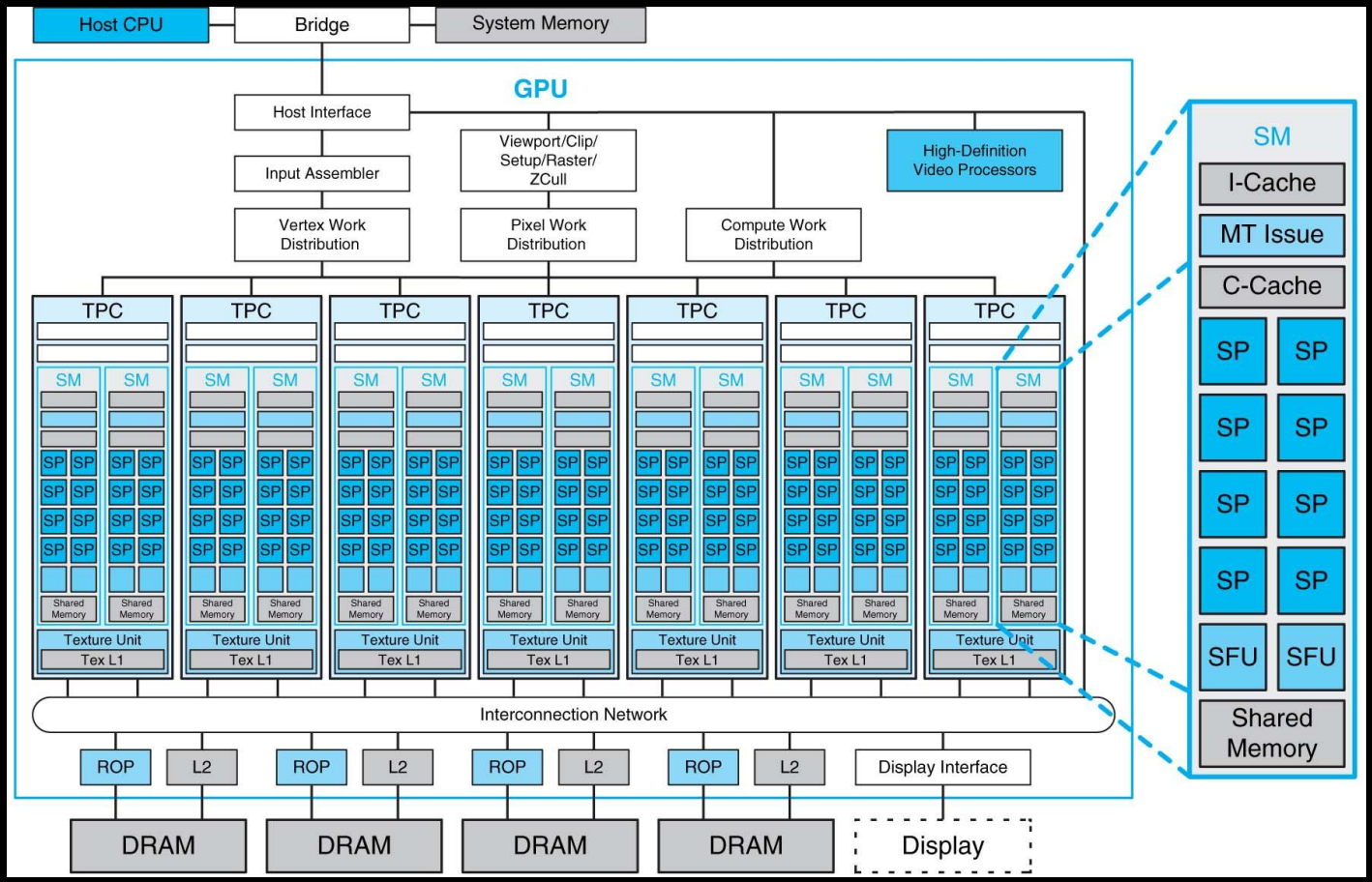


FIGURE A.2.5 Basic unified GPU architecture. Example GPU with 112 streaming processor (SP) cores organized in 14 streaming multiprocessors (SMs); the cores are highly multithreaded. It has the basic Tesla architecture of an NVIDIA GeForce 8800. The processors connect with four 64-bit-wide DRAM partitions via an interconnection network. Each SM has eight SP cores, two special function units (SFUs), instruction and constant caches, a multithreaded instruction unit, and a shared memory.

Can we use these for general computation?

Scientific Computing

- MATLAB codes

Convex hulls

Molecular Dynamics

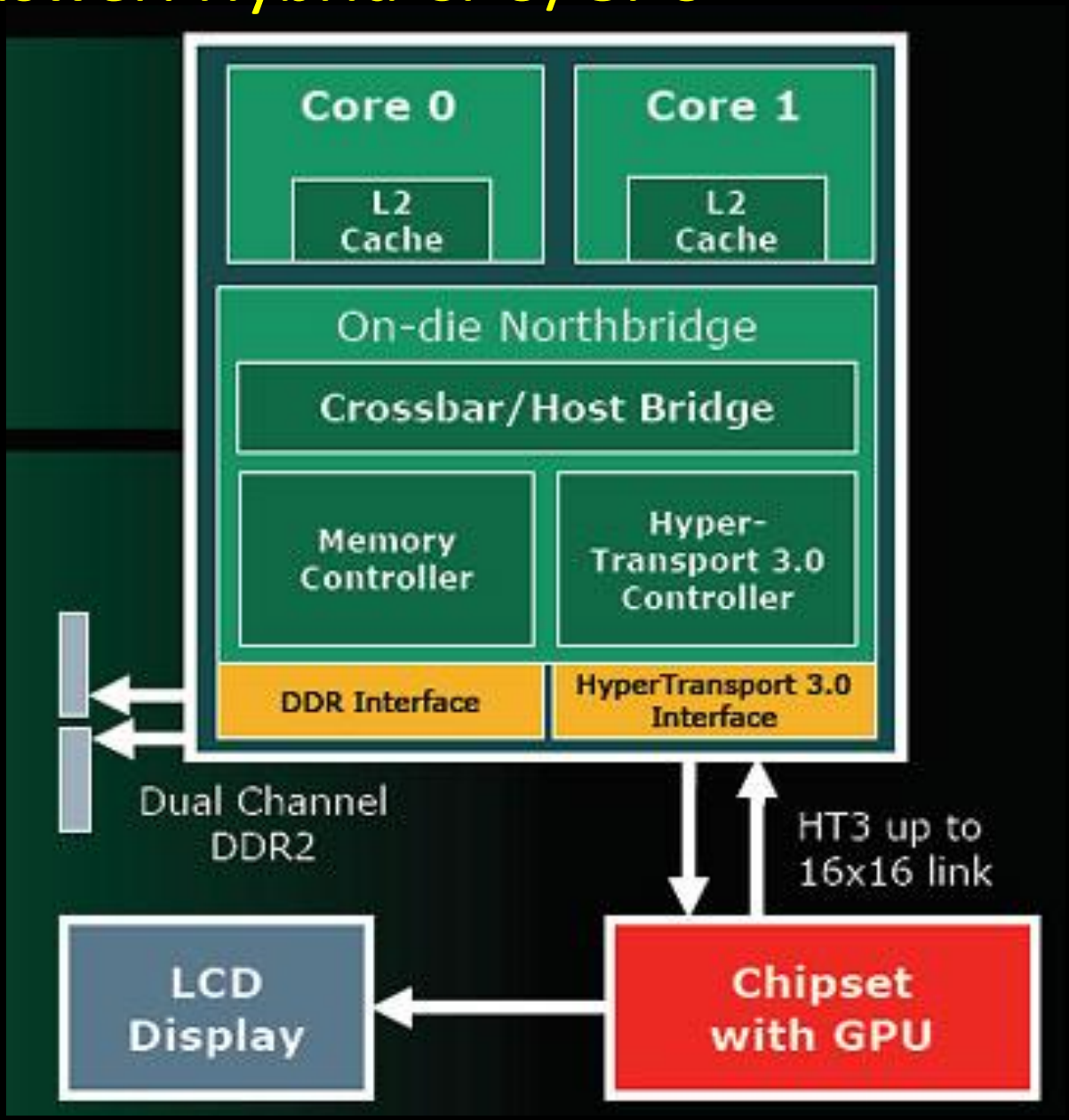
Etc.

NVIDIA's answer:

Compute Unified Device Architecture (CUDA)

- MATLAB/Fortran/etc. → "C for CUDA" → GPU Codes

AMD's Answer: Hybrid CPU/GPU

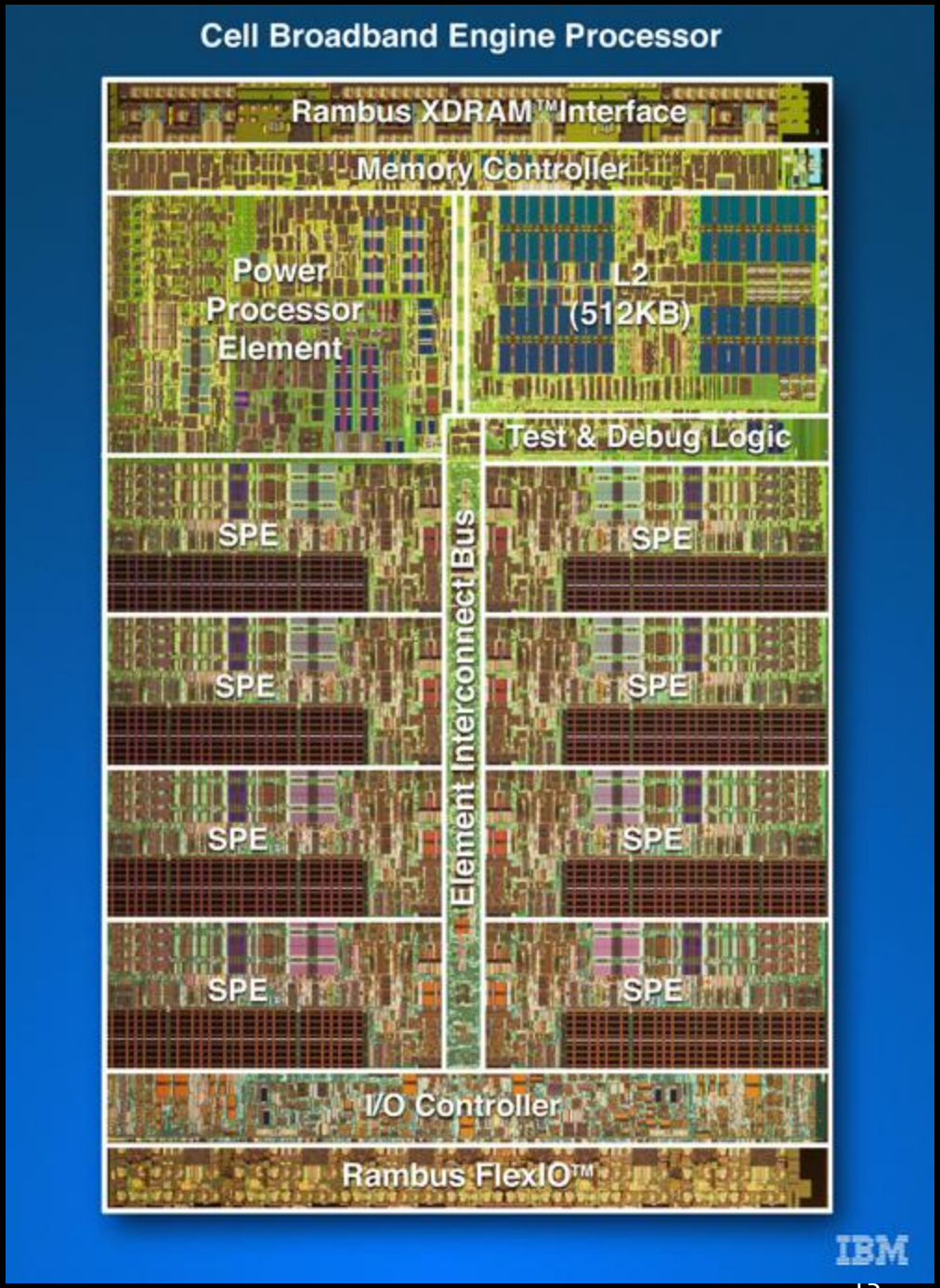


IBM/Sony/Toshiba

Sony Playstation 3

PPE

SPEs (synergestic)



Must exploit parallelism for performance

- Lots of parallelism in graphics applications
- Lots of parallelism in scientific computing

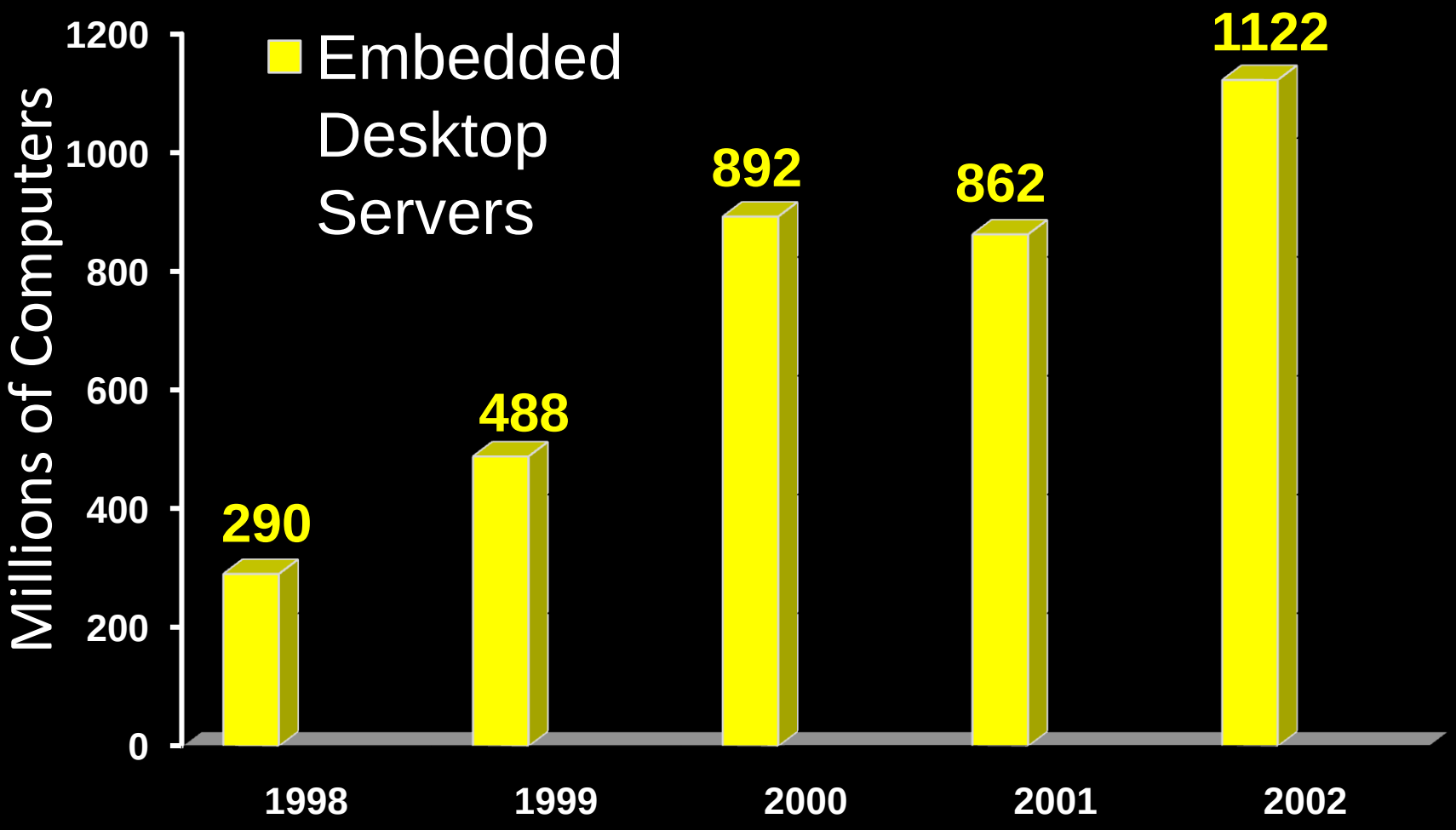
SIMD: single instruction, multiple data

- Perform same operation in parallel on many data items
- Data parallelism

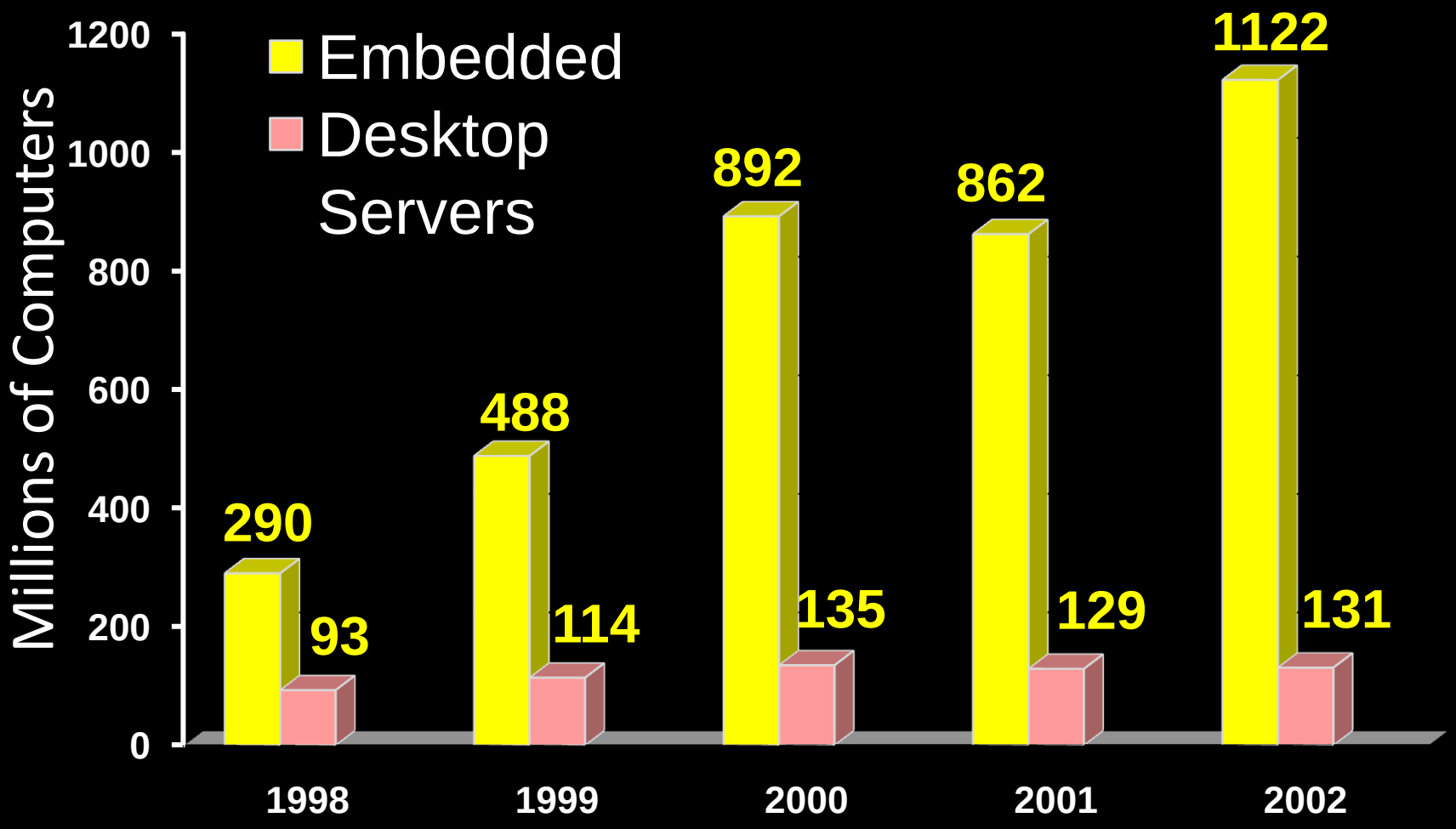
MIMD: multiple instruction, multiple data

- Run separate programs in parallel (on different data)
- Task parallelism

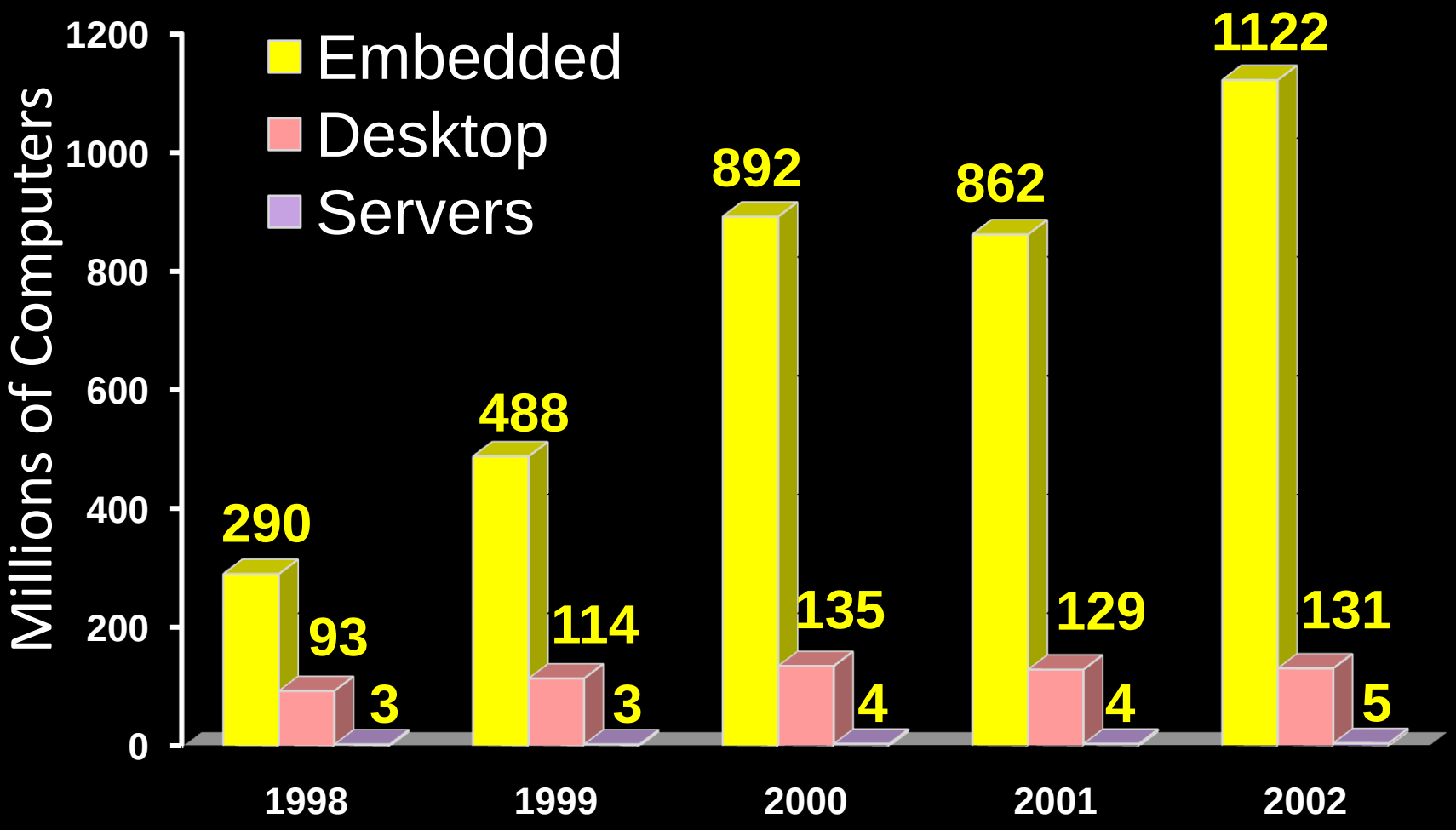
Where is the Market?



Where is the Market?



Where is the Market?




151st St to US-69 S
KIOWA
ARROWHEAD
LINDENWOOD
Arrival
2:16
Menu
Turn In
6.0

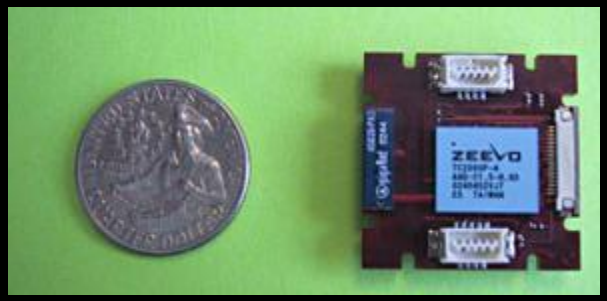

palm
vodafone
Pictures & Videos
Send
Menu
Treo

Smart Dust....


WASP MAV

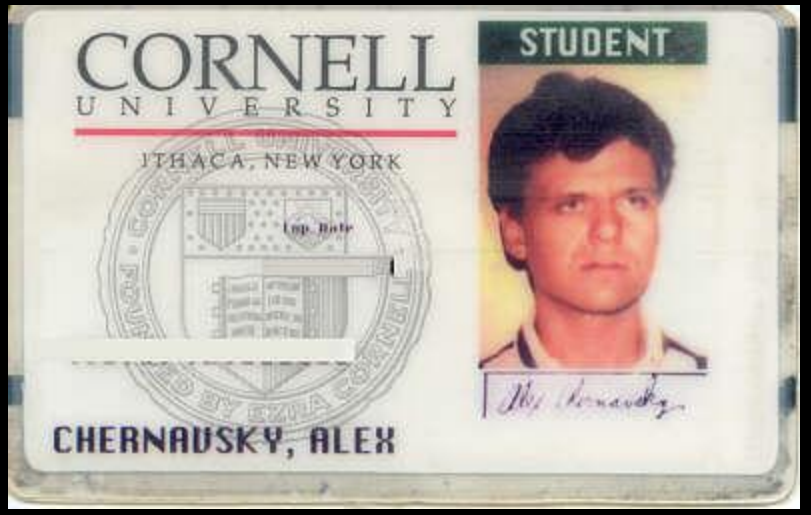

Sensor
Network Node
Network Node


Network Node

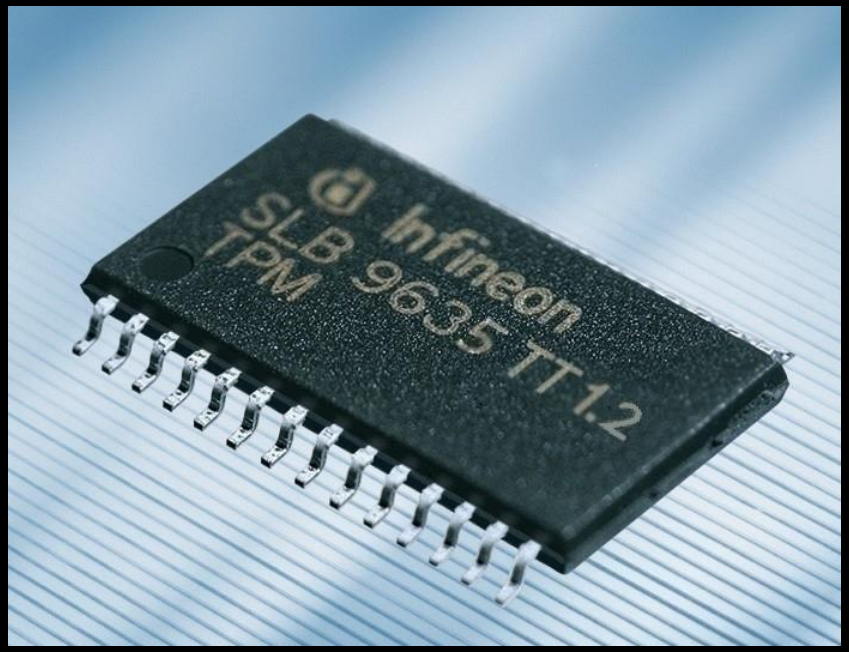

ZEEVO

Smart Cards…


carte d'assurance maladie
vitale

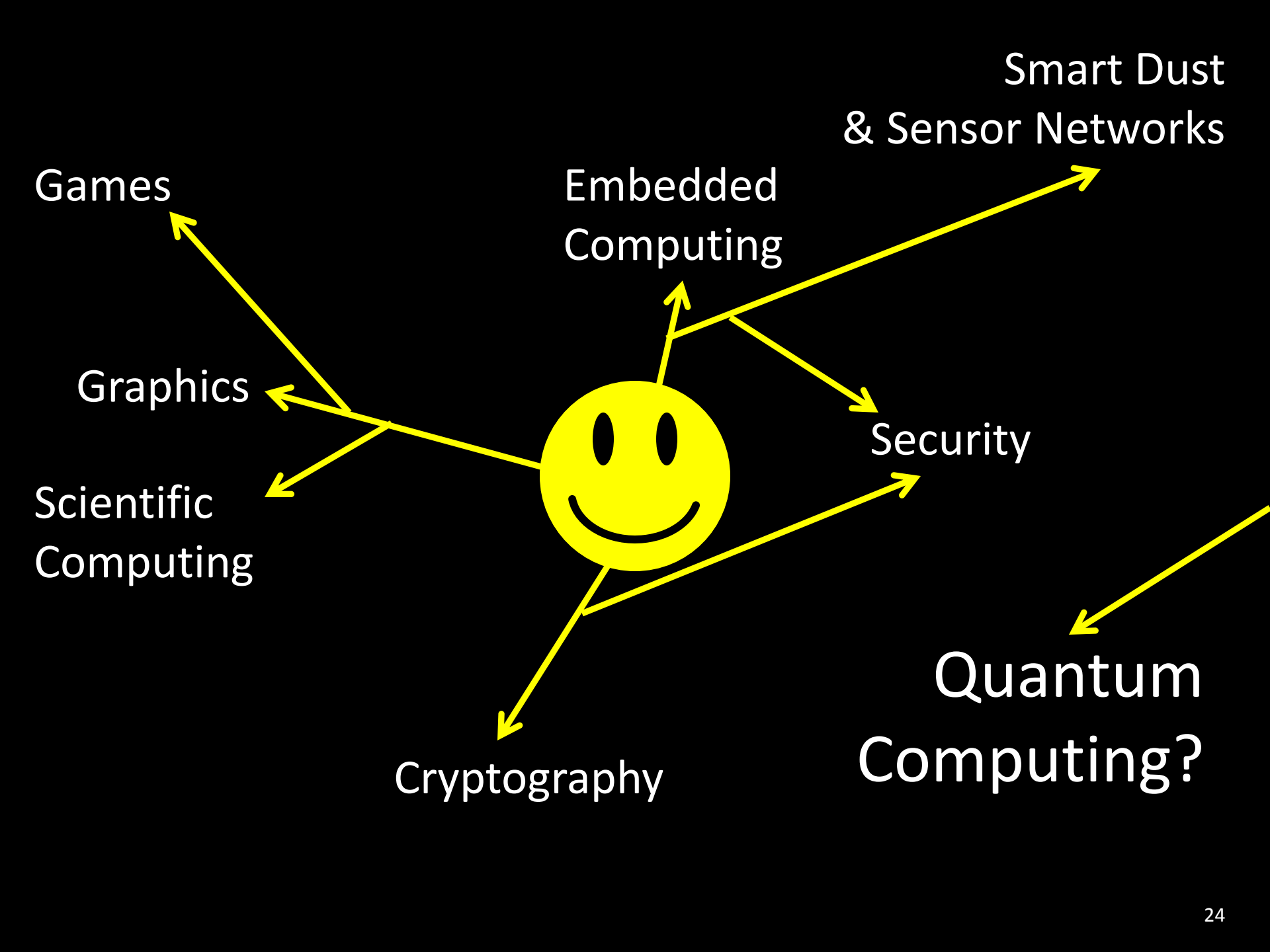

CORNELL
UNIVERSITY
ITHACA, NEW YORK
STUDENT
CHERNAUSKY, ALEX

Cryptography and security…



TPM 1.2

IBM 4758
Secure Cryptoprocessor


Smart Dust
& Sensor Networks
Games
Embedded
Computing
Graphics
Security
Scientific
Computing
Cryptography
Quantum
Computing?

How useful is this class, in all seriousness, for a computer scientist going into software engineering, meaning not low-level stuff?

How much of computer architecture do software engineers actually have to deal with?

What are the most important aspects of computer architecture that a software engineer should keep in mind while programming?

These days, programs run on hardware...

… more than ever before

Google Chrome

→ Operating Systems

→ Multi-Core & Hyper-Threading

→ Datapath Pipelines, Caches, MMUs, I/O & DMA

→ Busses, Logic, & State machines

→ Gates

→ Transistors

→ Silicon

→ Electrons

CS 3110: Better concurrent programming

CS 4410: The Operating System!

CS 4450: Networking

CS 4620: Graphics

~~CS 4821: Quantum Computing~~

And many more…

Thank you!

If you want to make an apple pie from scratch, you must first create the universe.

– Carl Sagan