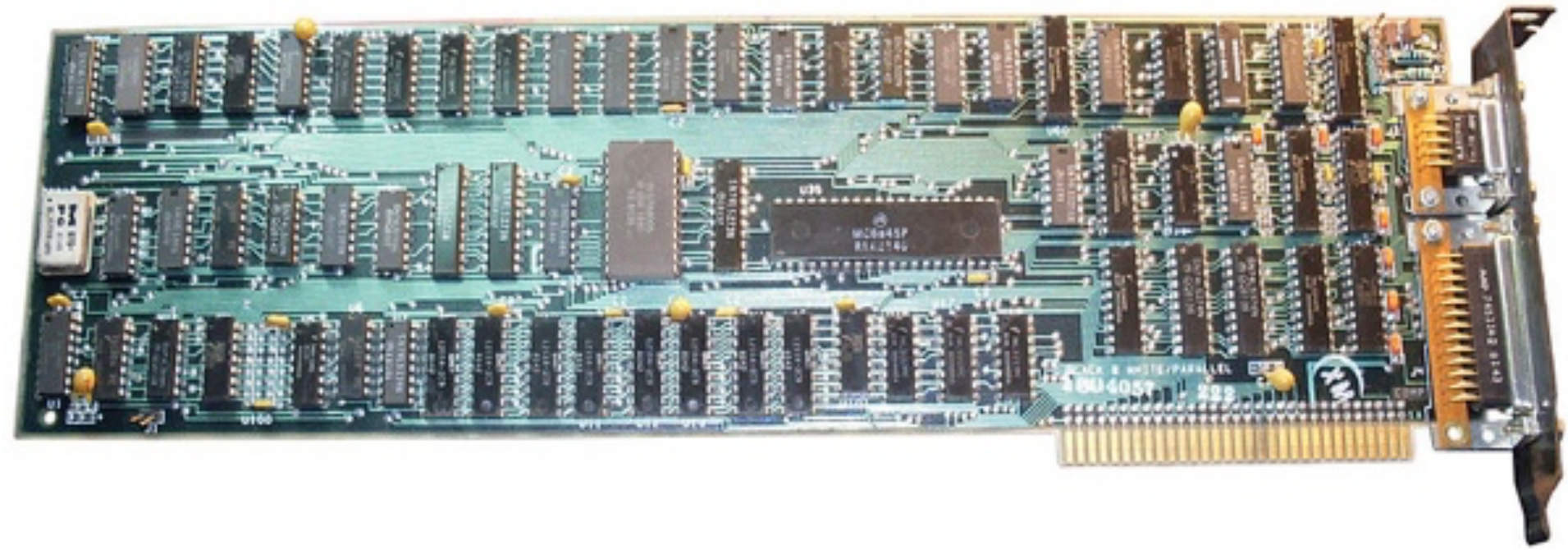
- Museum (this Wednesday)
- No class *next* Wednesday
- Data center visit on the 24th (more details to come)
- Midterm II May 31st
- 2 more class days :(

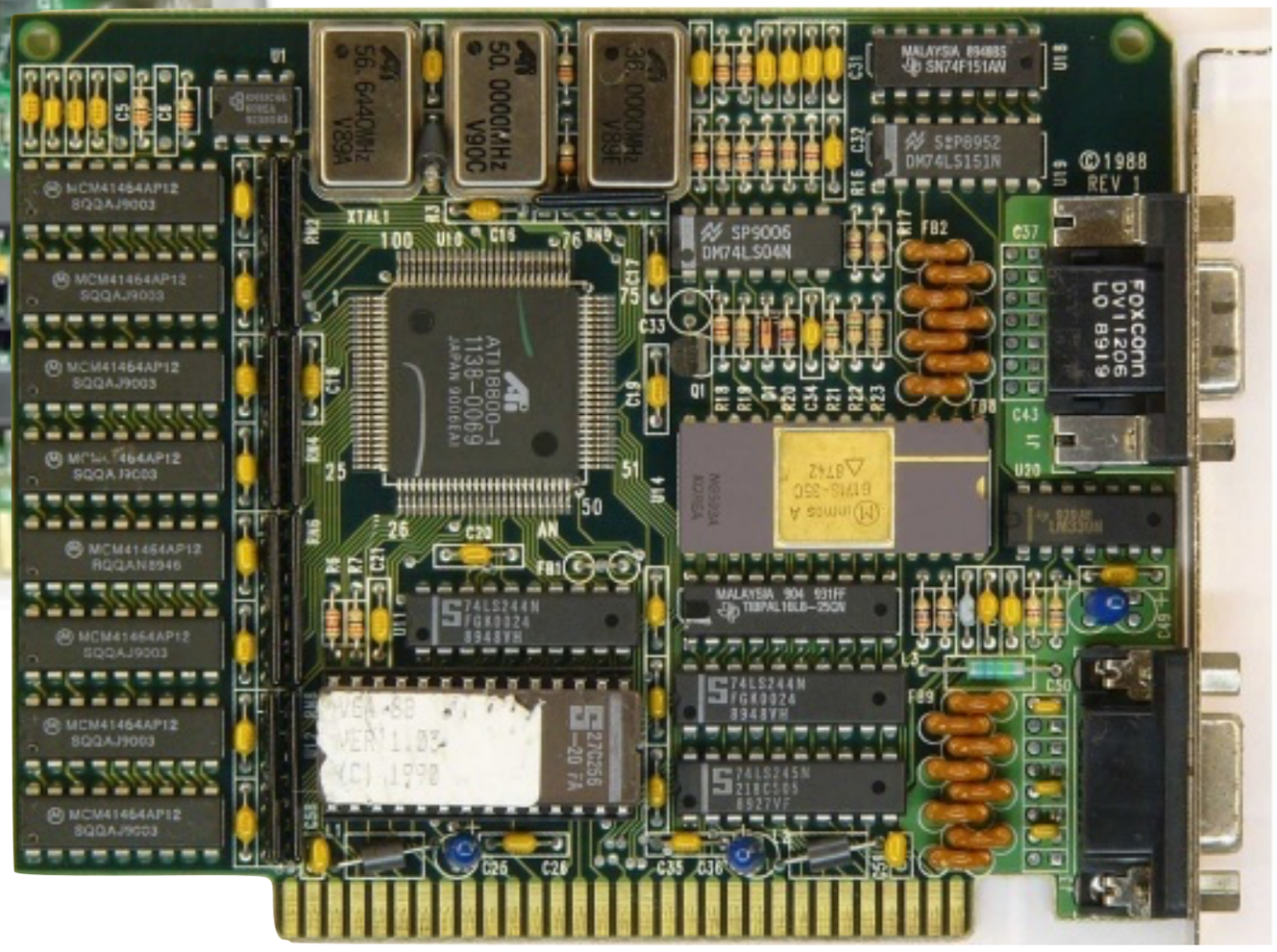
GPUs



IBM monochrome adapter (1981)
w/parallel printer port :)

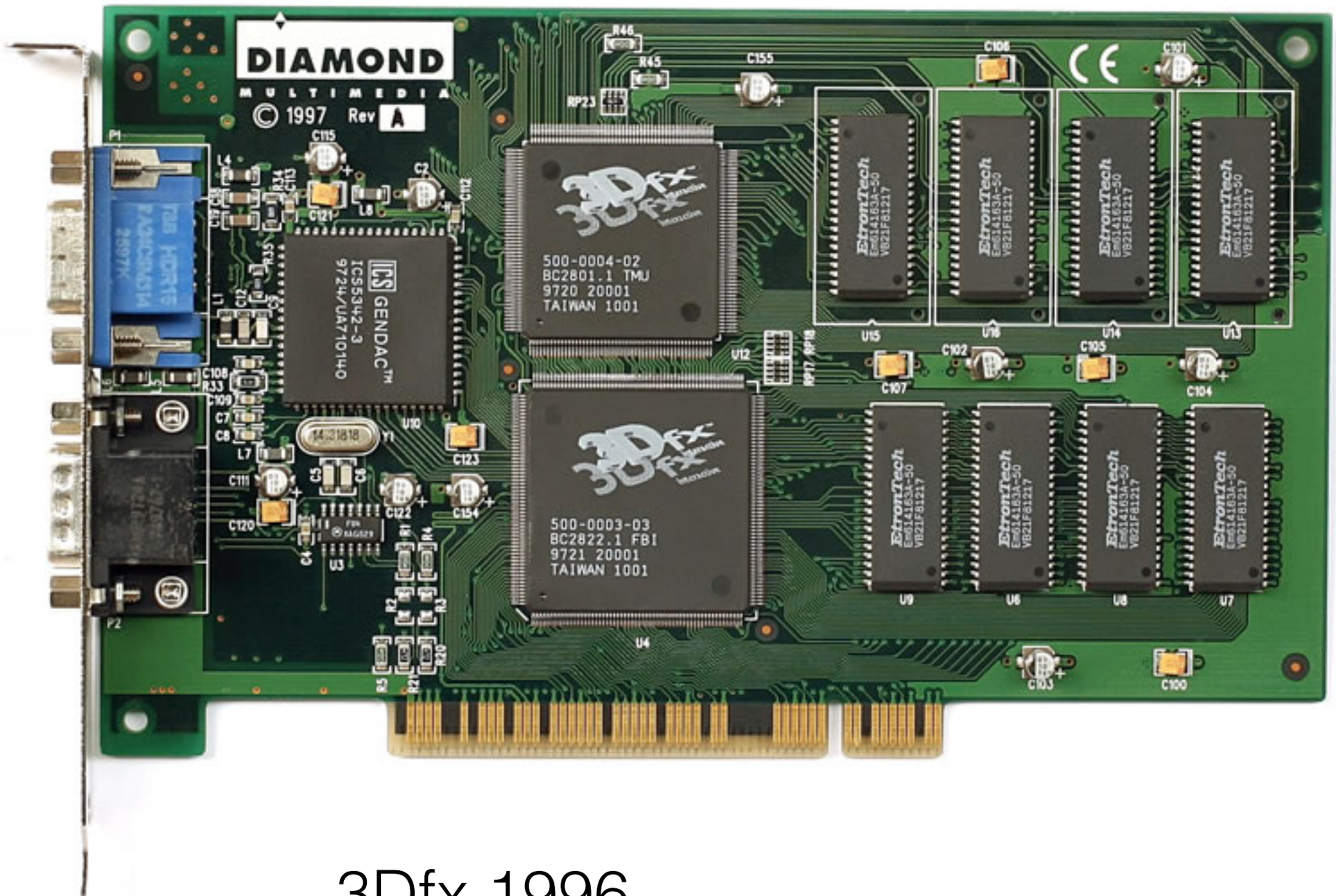
Characters and raster mode

ATI (now AMD) EGA Wonder (1987)
2D Graphics

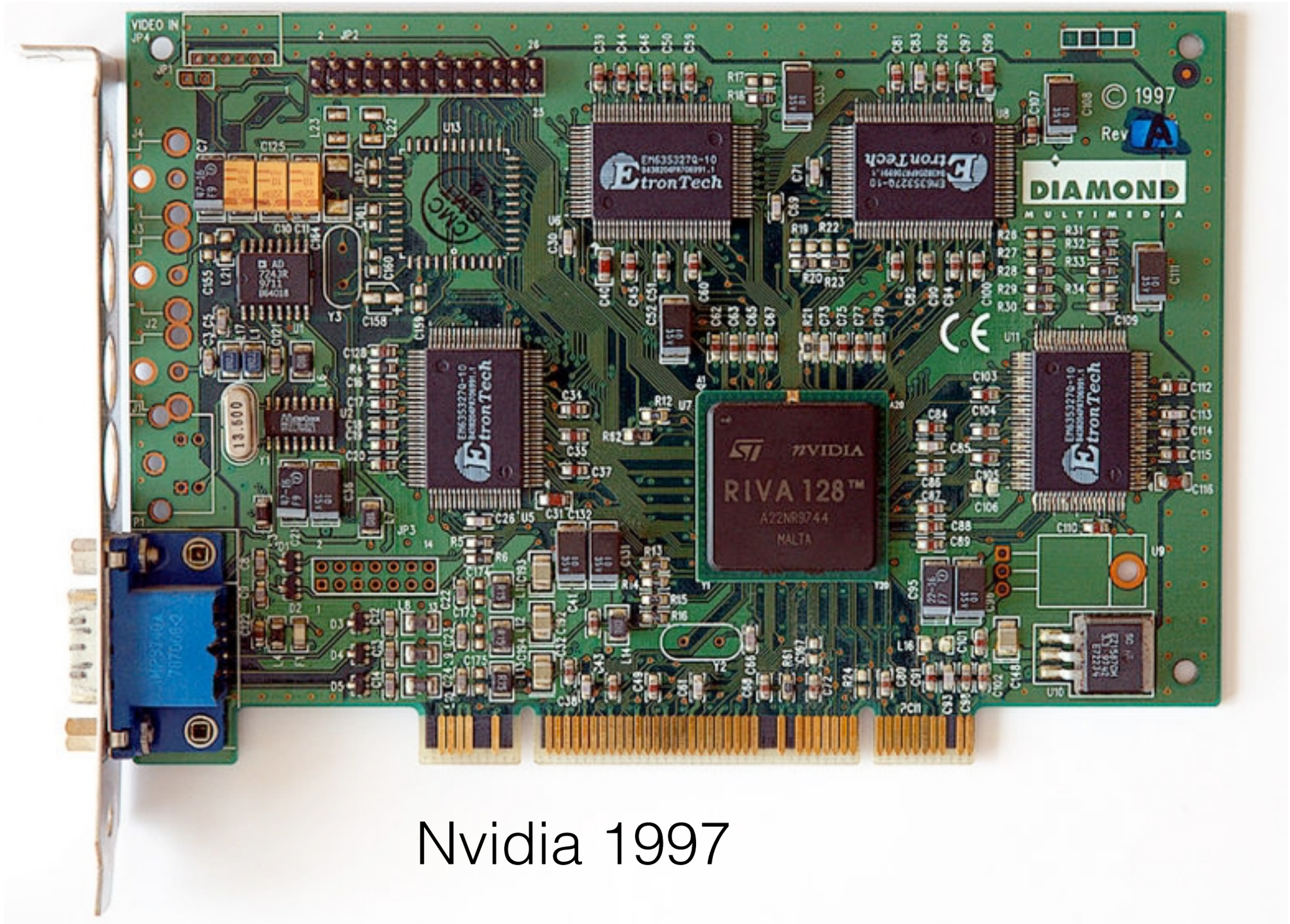


ATI VGA Wonder 1988

S3 Trio (1995) 3D Graphics


DIAMOND
MULTIMEDIA
© 1997 Rev A
500-0004-02
BC2801.1 TMU
9720 20001
TAIWAN 1001
500-0003-03
BC2822.1 FBI
9721 20001
TAIWAN 1001
ICS GENDAC™
ICS5342-3
9724/UA710140
EtronTech
EM6141163A-50
VB21F81217

3Dfx 1996

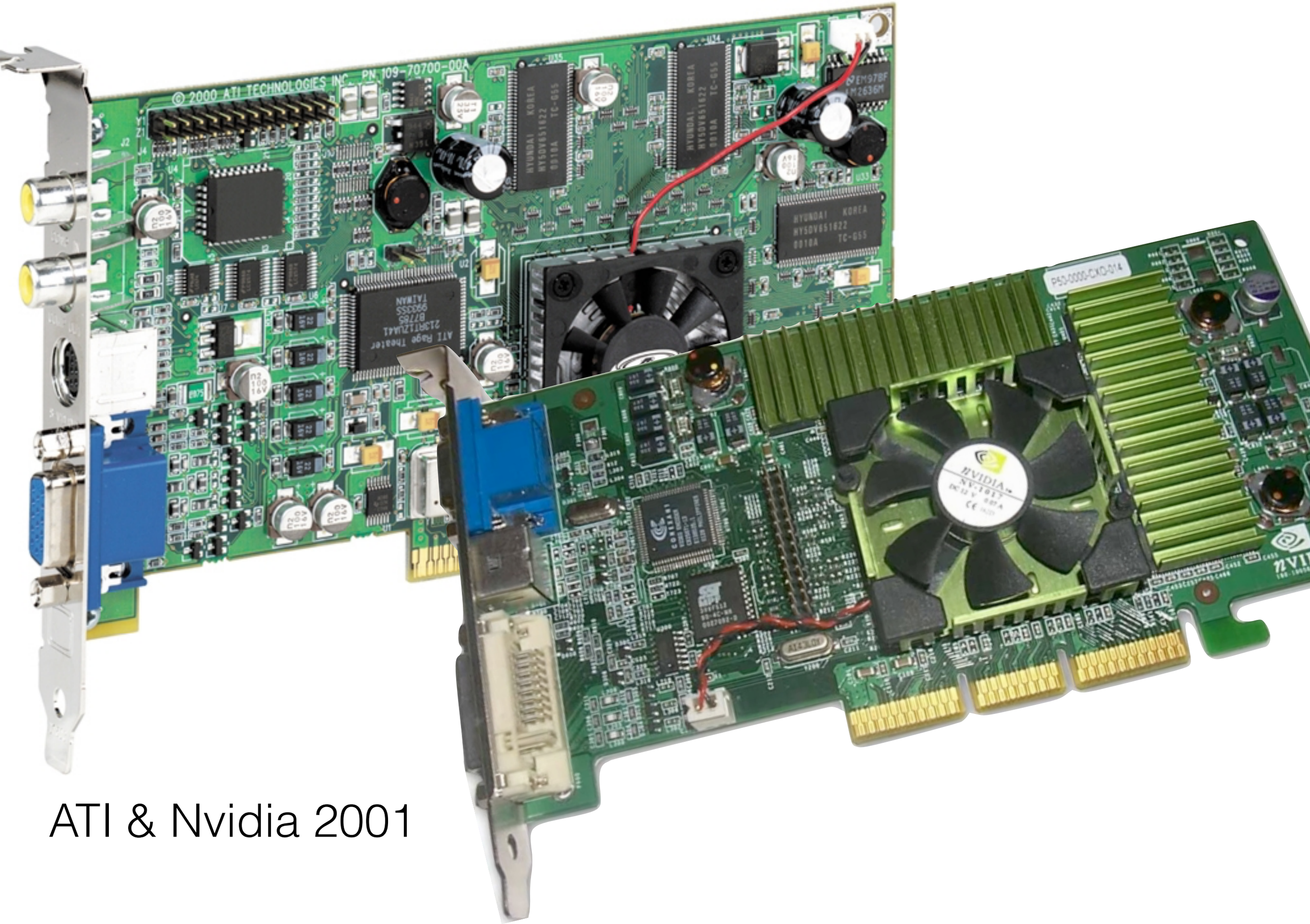

VIDEO IN
JP4
JP2
© 1997
Rev A
DIAMOND
MULTIMEDIA
EtronTech
EM6353270-10
NVIDIA
RIVA 128™
A22NR9744
MALTA
13.500

Nvidia 1997

1998 *Note fans!*

3Dfx Voodoo 5 1999

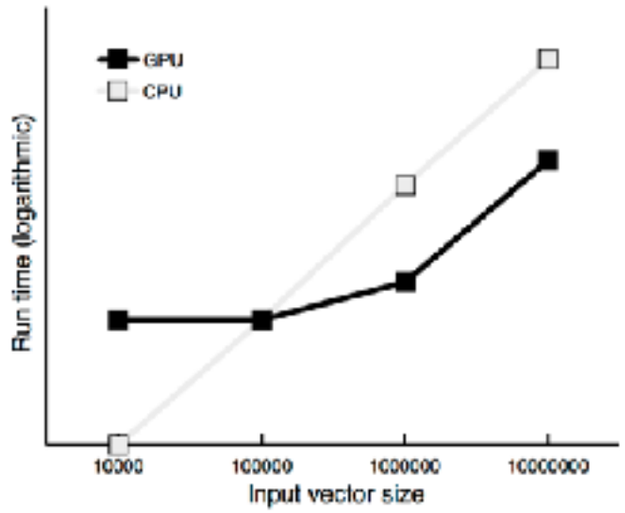
Note fans & auxiliary power connector



ATI & Nvidia 2001

Using Modern Graphics Architectures for General-Purpose Computing: A Framework and Analysis

Chris J. Thompson Sahngyun Hahn Mark Oskin
Department of Computer Science and Engineering
University of Washington
{*cthomp, syhahn, oskin*}*@cs.washington.edu*



Winter 2002, published Fall 2002

Table 3. The vertex program instruction set.

Opcode	Description
ARL	Address register load
MOV	Move
MUL	Multiply
ADD	Add
SUB	Subtract
MAD	Multiply and add
ABS	Absolute value
RCP	Reciprocal
RCC	Reciprocal (clamped)
RSQ	Reciprocal square root
DP3	3-component dot product
DP4	4-component dot product
DPH	Homogenous dot product
DST	Cartesian distance
MIN	Minimum
MAX	Maximum
SLT	Set on less than
SGE	Set on greater/equal than
EXP	Exponential base 2
LOG	Logarithm base 2
LIT	Light coefficient formula

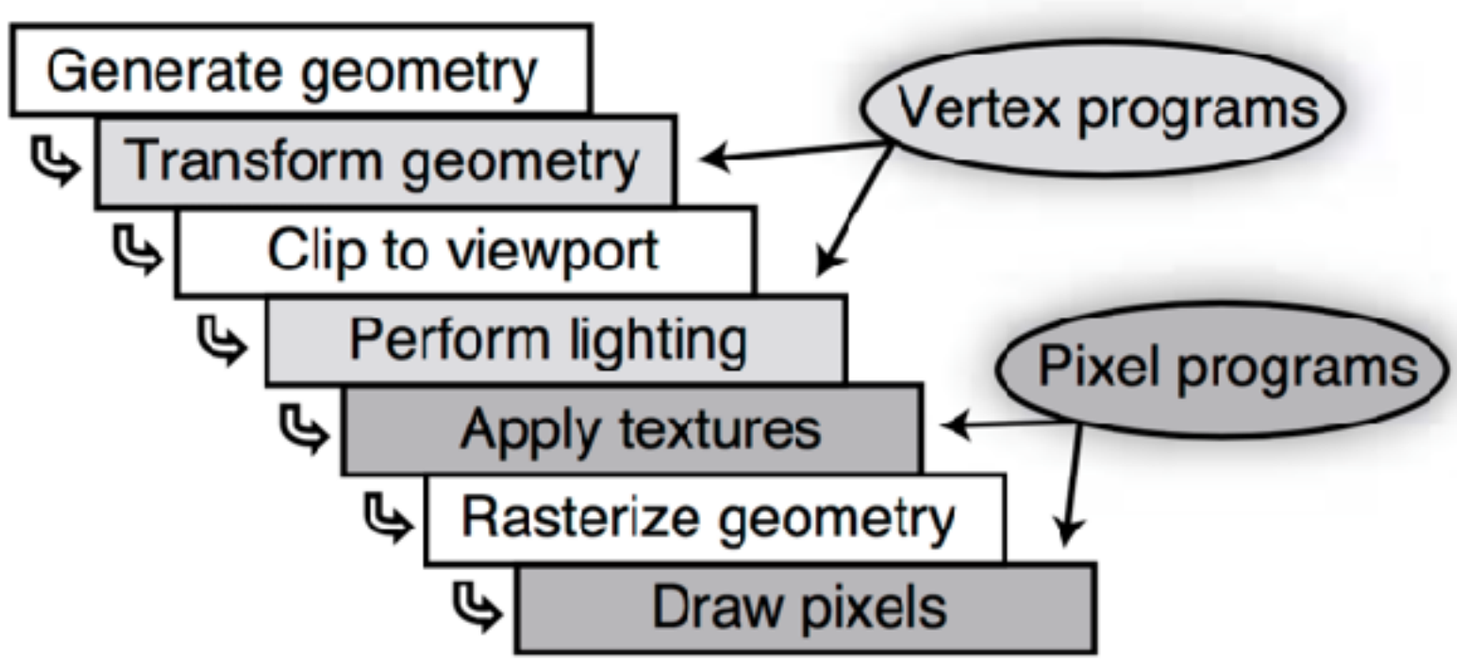
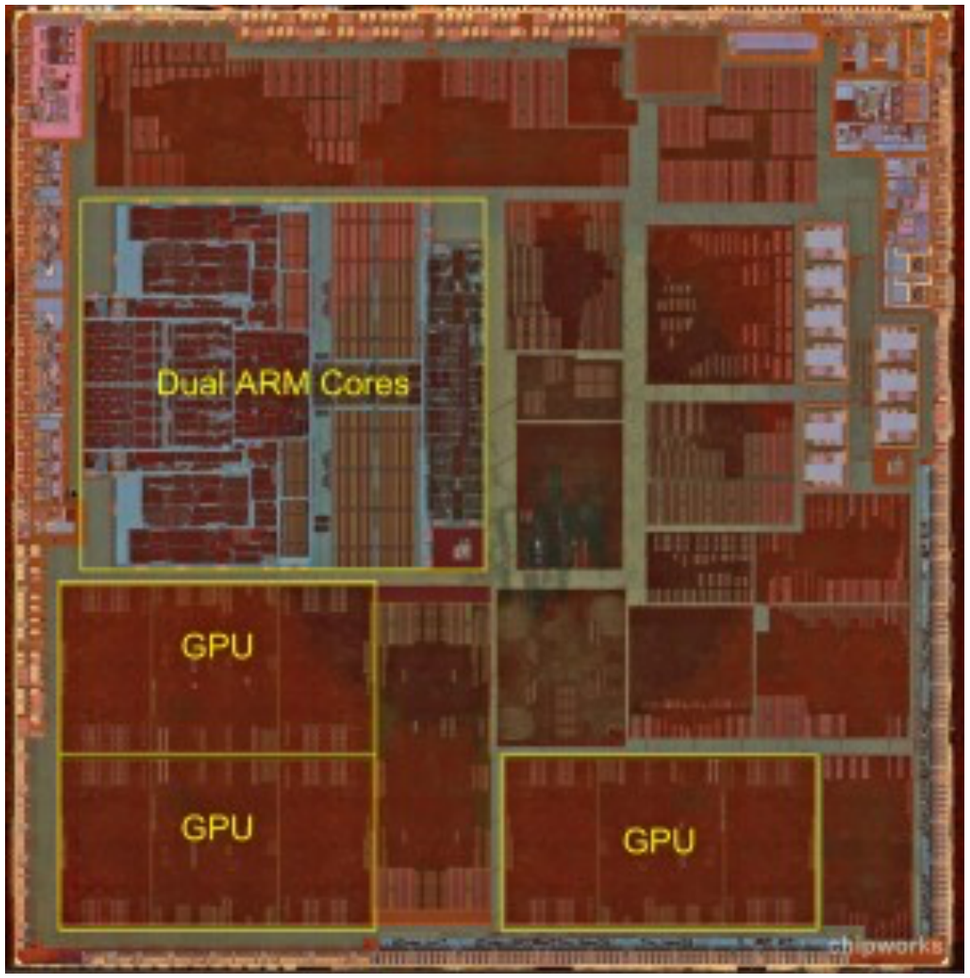


Figure 2. A programmable graphics pipeline.

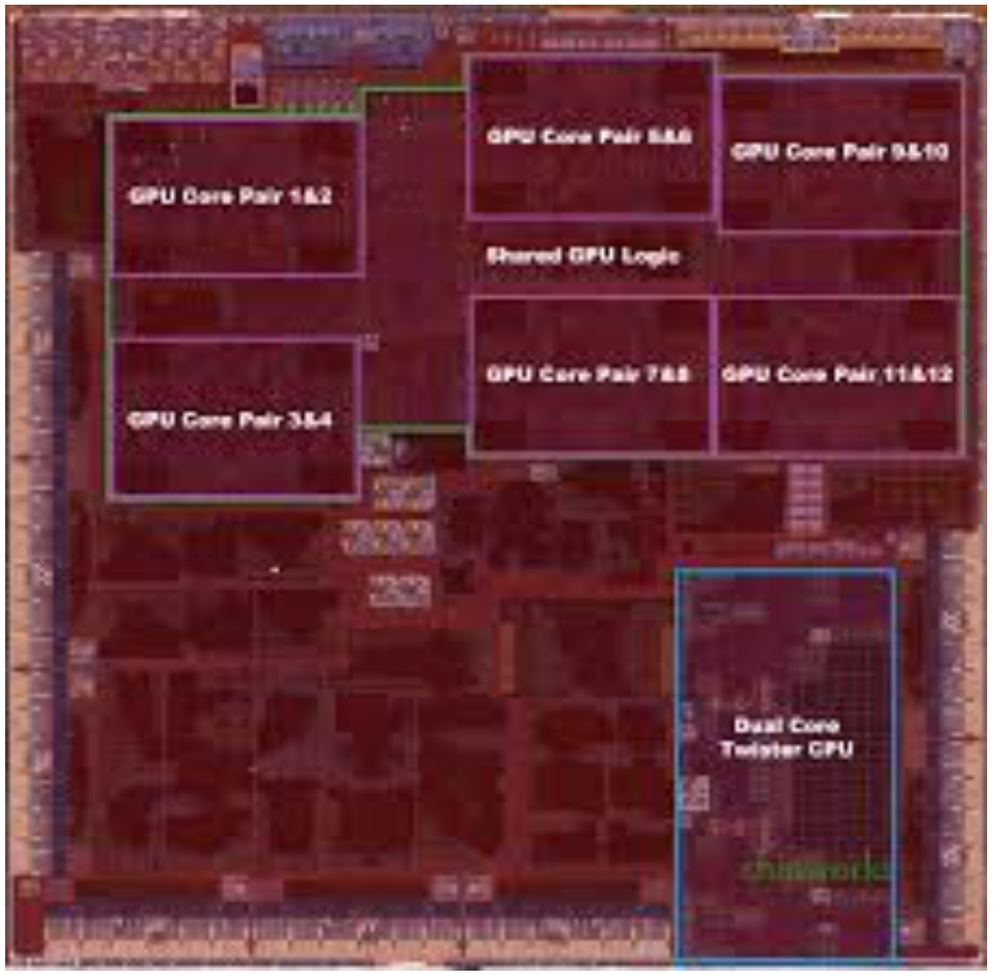
2005

2007

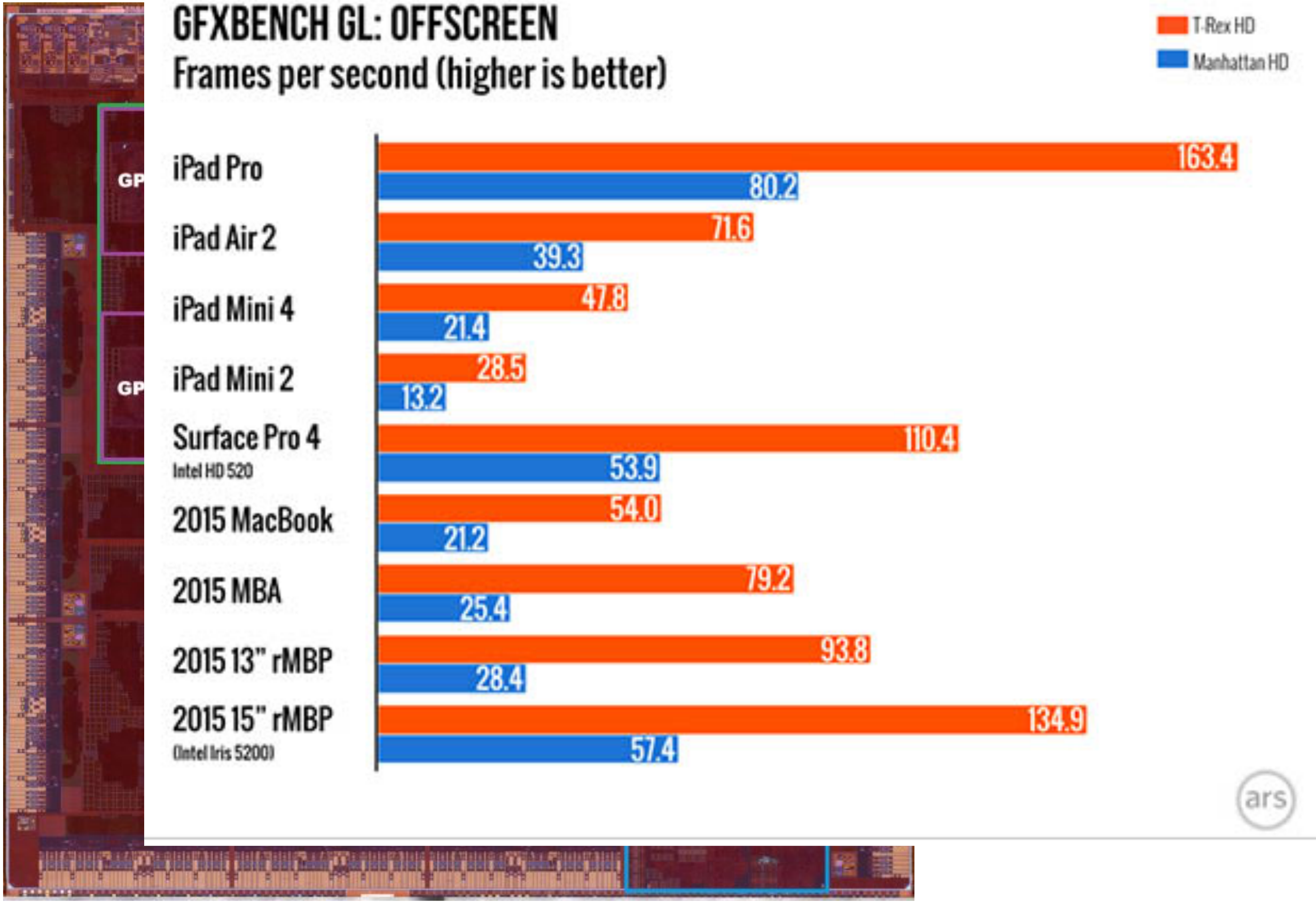
2009


Dual ARM Cores
GPU
GPU
GPU
chipworks

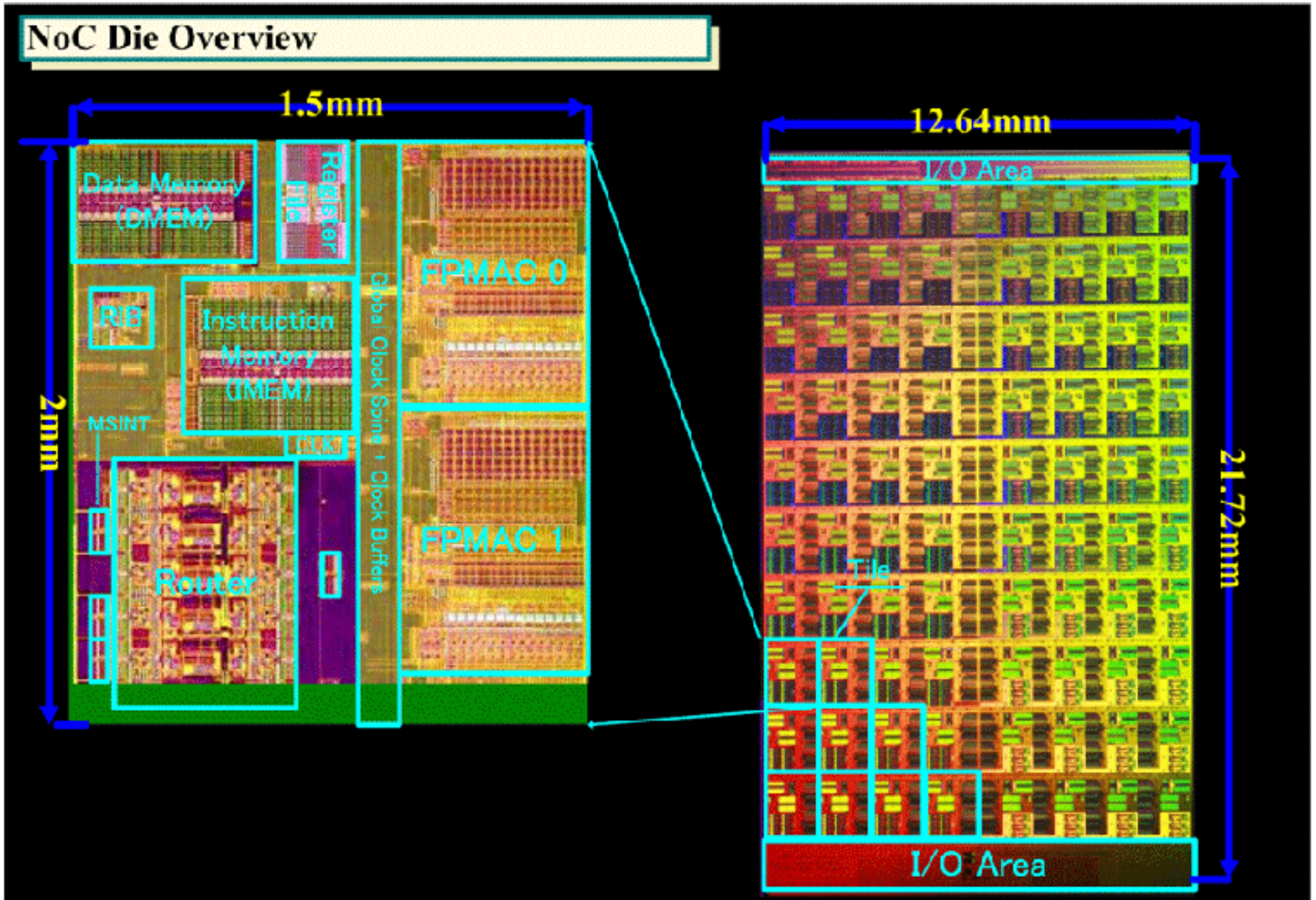
A6


GPU Core Pair 1&2
GPU Core Pair 3&4
GPU Core Pair 5&6
GPU Core Pair 9&10
Shared GPU Logic
GPU Core Pair 7&8
GPU Core Pair 11&12
Dual Core
Twister CPU
chipworks

A9X

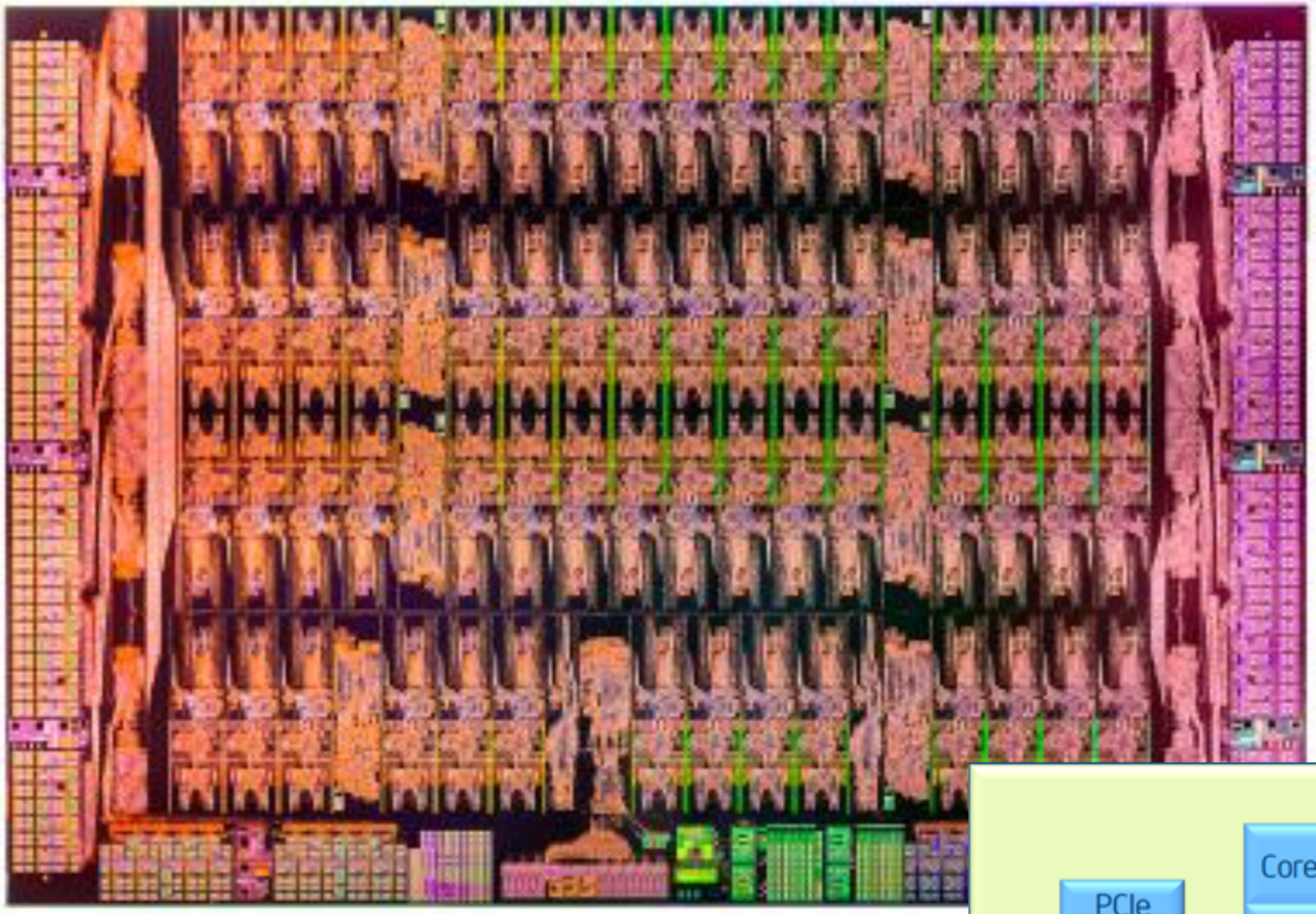

GFXBENCH GL: OFFSCREEN
Frames per second (higher is better)
T-Rex HD
Manhattan HD
iPad Pro
163.4
80.2
iPad Air 2
71.6
39.3
iPad Mini 4
47.8
21.4
iPad Mini 2
28.5
13.2
Surface Pro 4
Intel HD 520
110.4
53.9
2015 MacBook
54.0
21.2
2015 MBA
79.2
25.4
2015 13" rMBP
93.8
28.4
2015 15" rMBP
(Intel Iris 5200)
134.9
57.4
ars

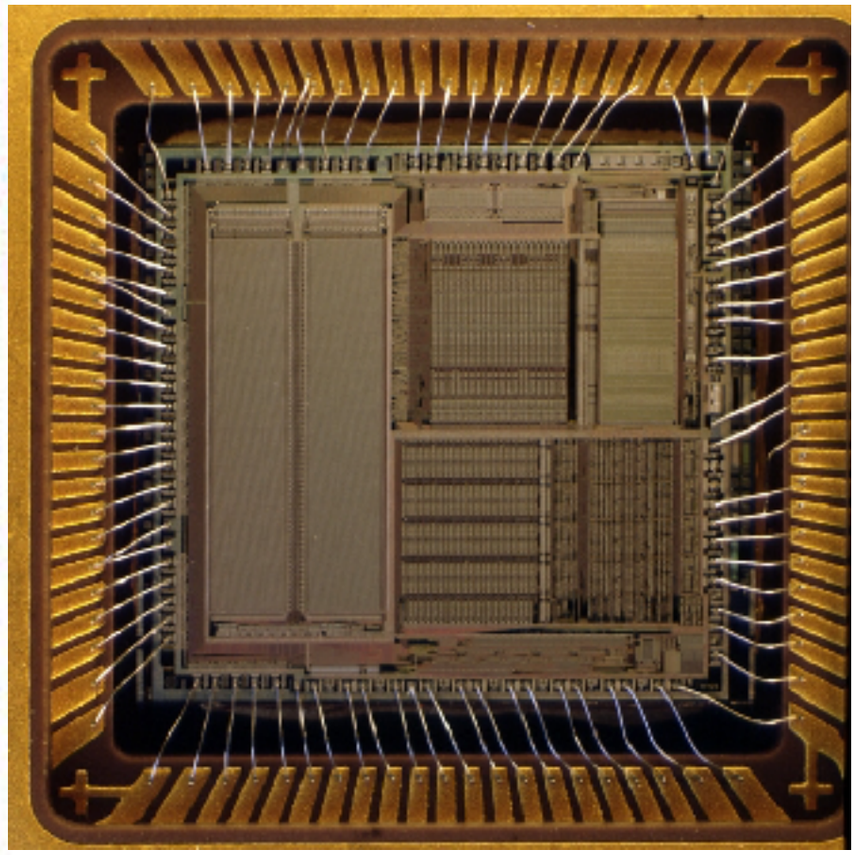
Apple A9X



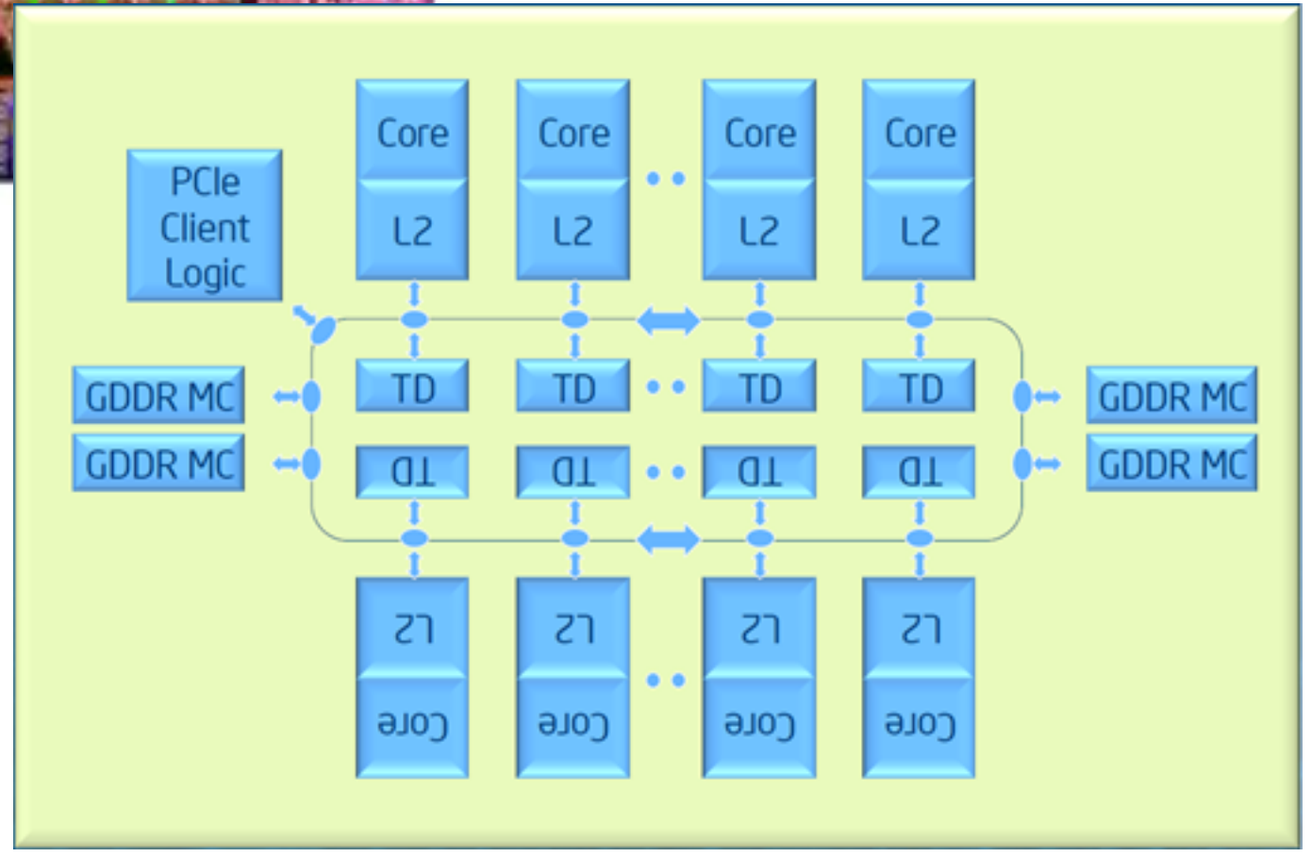
Tile	
Tile Transistors	1.2 M Transistors
Tile Area	3mm2 (1.5mm x 2mm)
Router Transistors	210k Transistors
Router Area	0.34mm2

Full chip	
Process Technology	65nm CMOS
Interconneet	1 poly, 8 metal(Cu)
Transistors	100 M Transistors
Die Area	275mm2 (12.64mm x 21.72mm)
C4 bumps	8390
Package	1248 pin LGA, 14 layers, 343 signal pins





(*the poorly named)* Intel MIC (2010)



What are the key enabling technologies behind GPUs?

- Programmable pipeline
- Abstract ISA / API
- It's all about the memory
- High bandwidth PCIe is key for GPGPU
- Why can they use SIMD?
  - it is data parallel computation
  - the control flow is largely the same
- Need a lot of parallel tasks

Figure 3: GCN Compute Unit

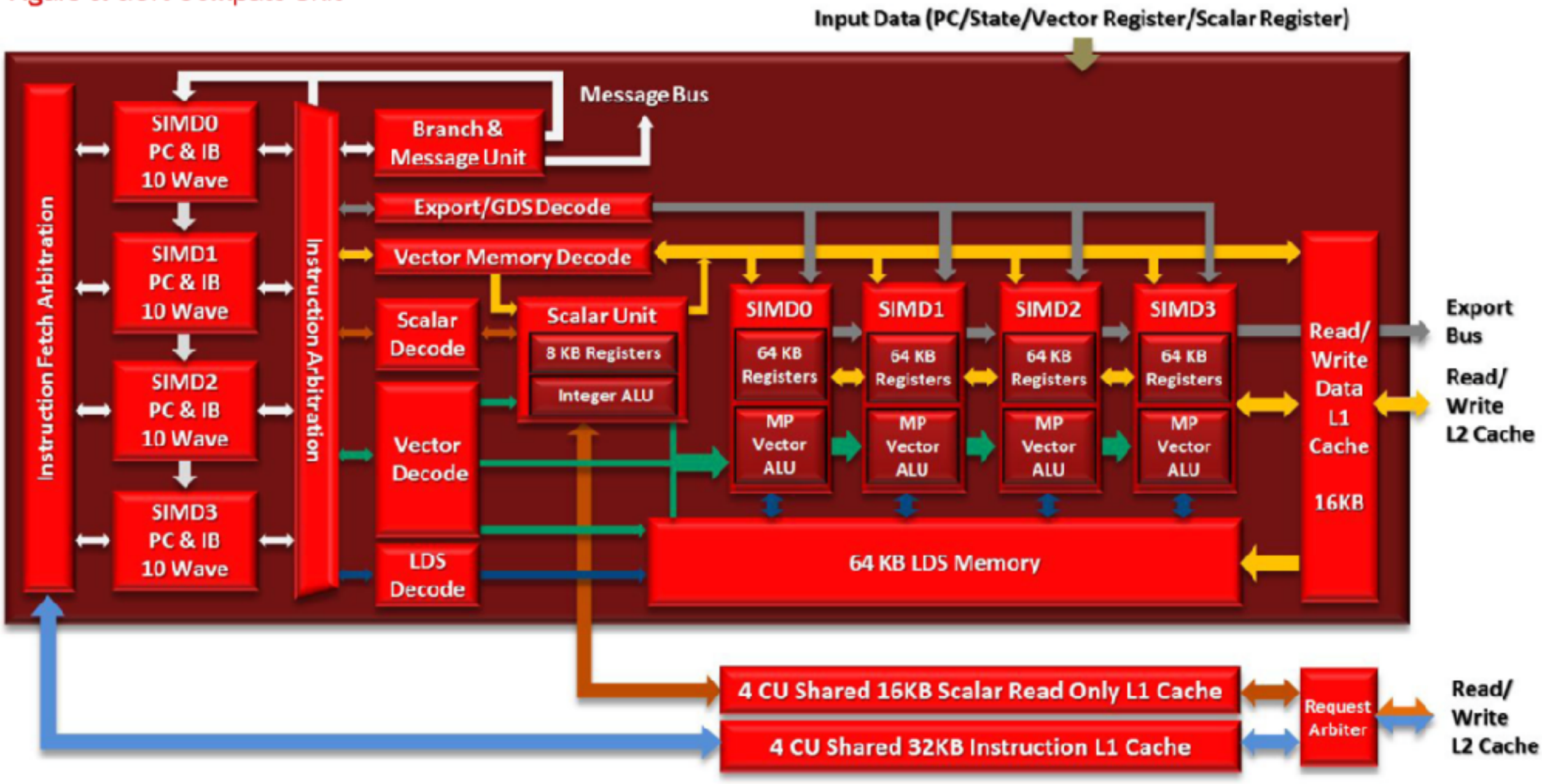


Figure 4: Local Data Share (LDS)

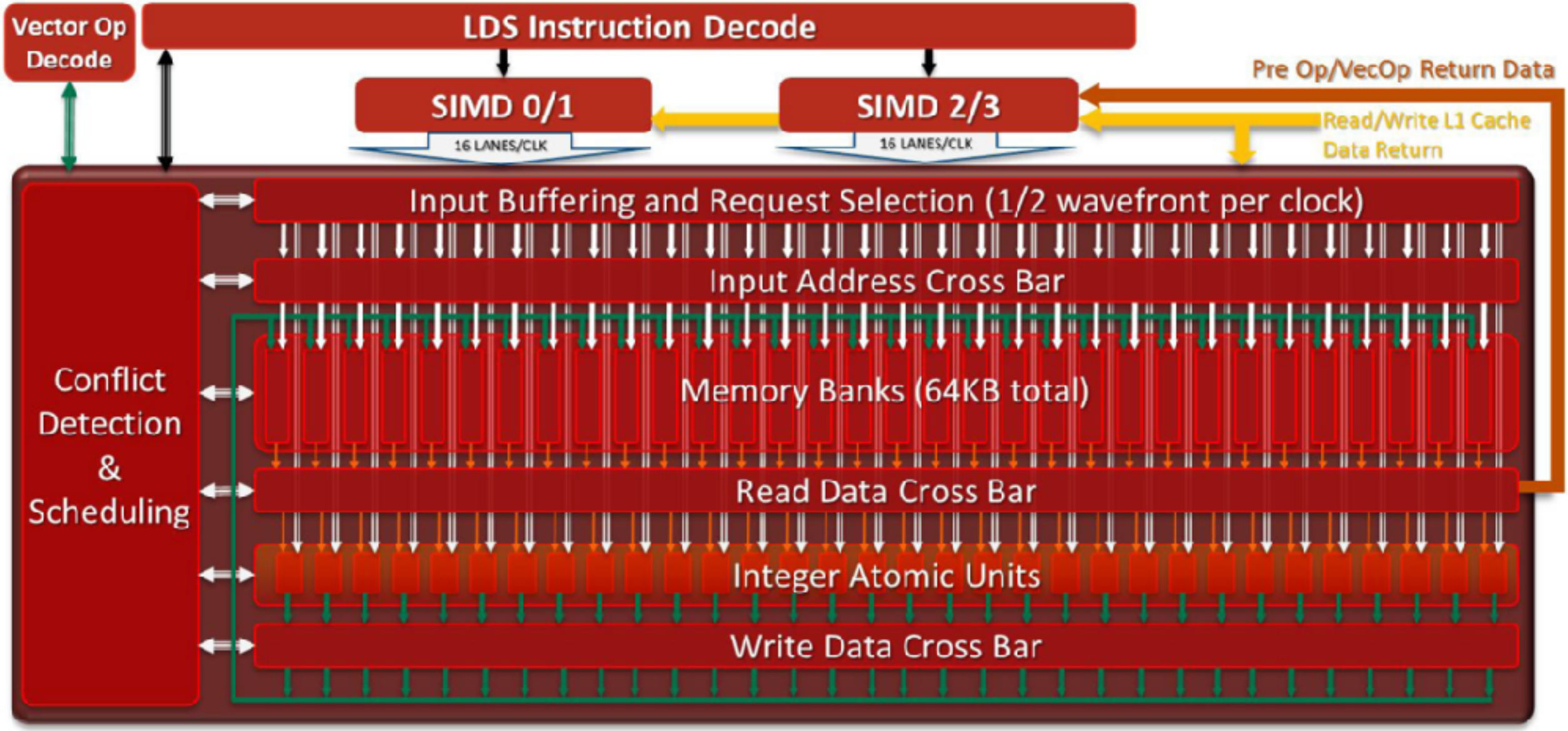


Figure 6: Cache Hierarchy

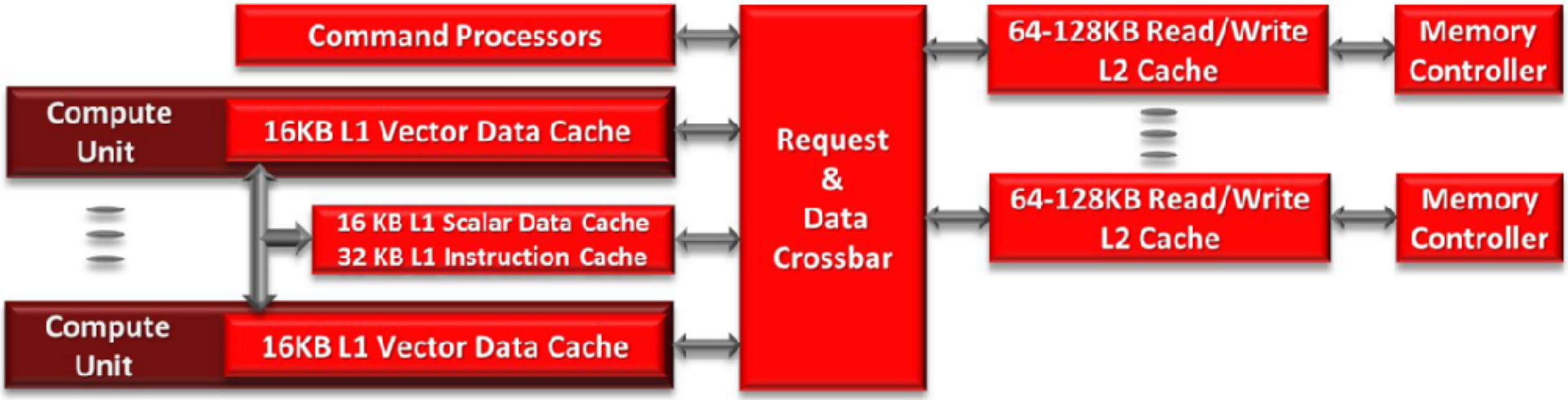
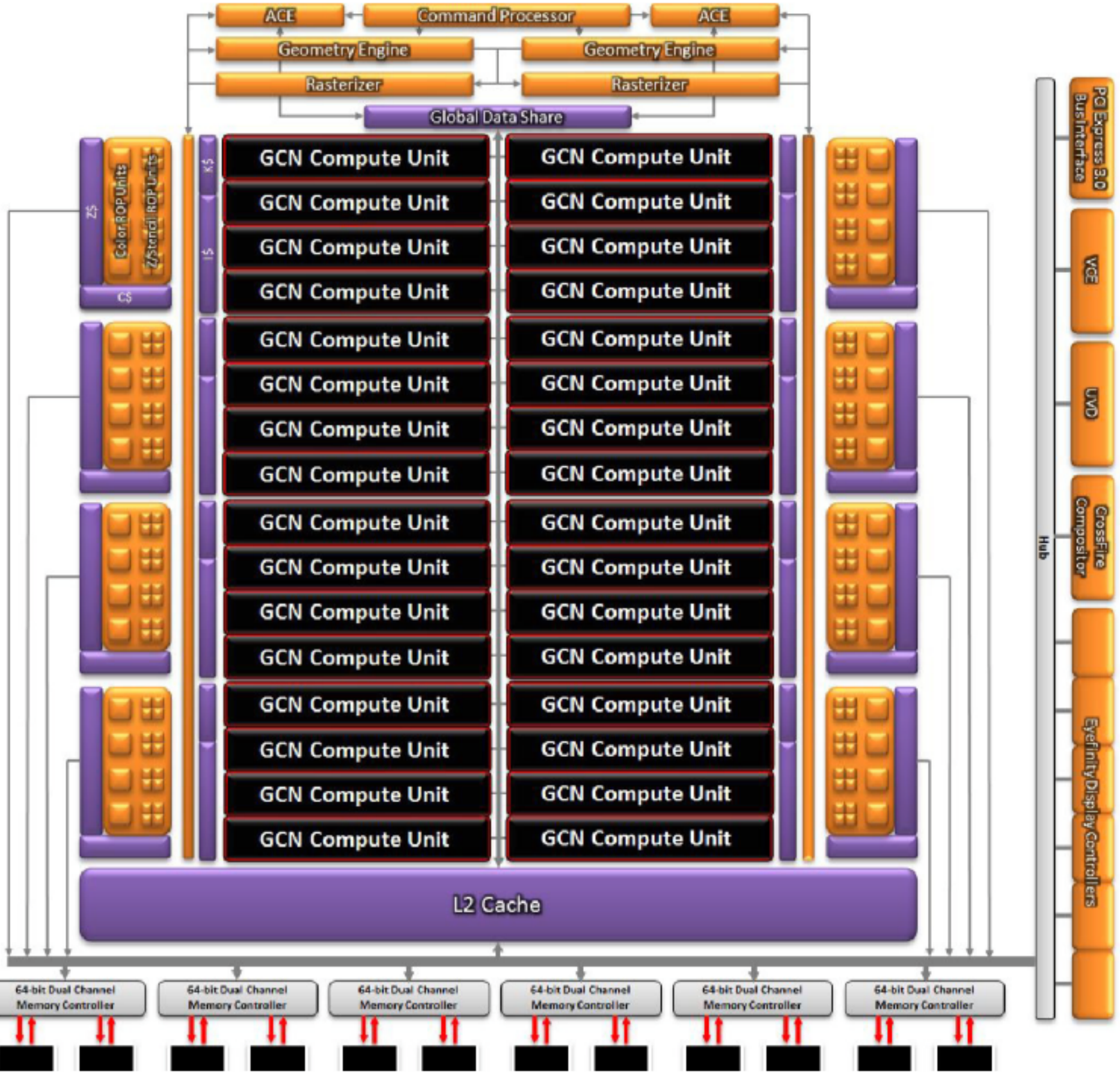
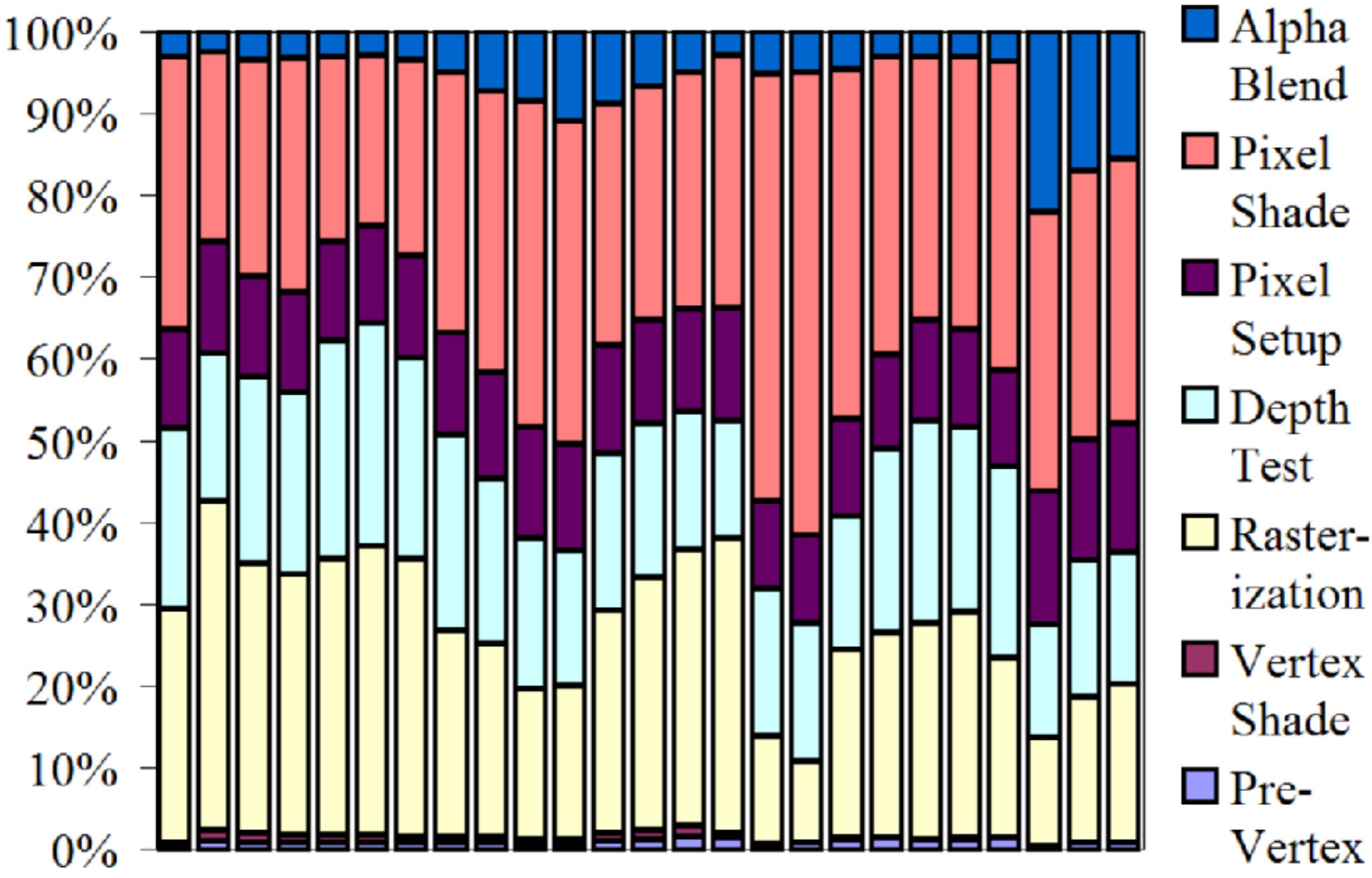
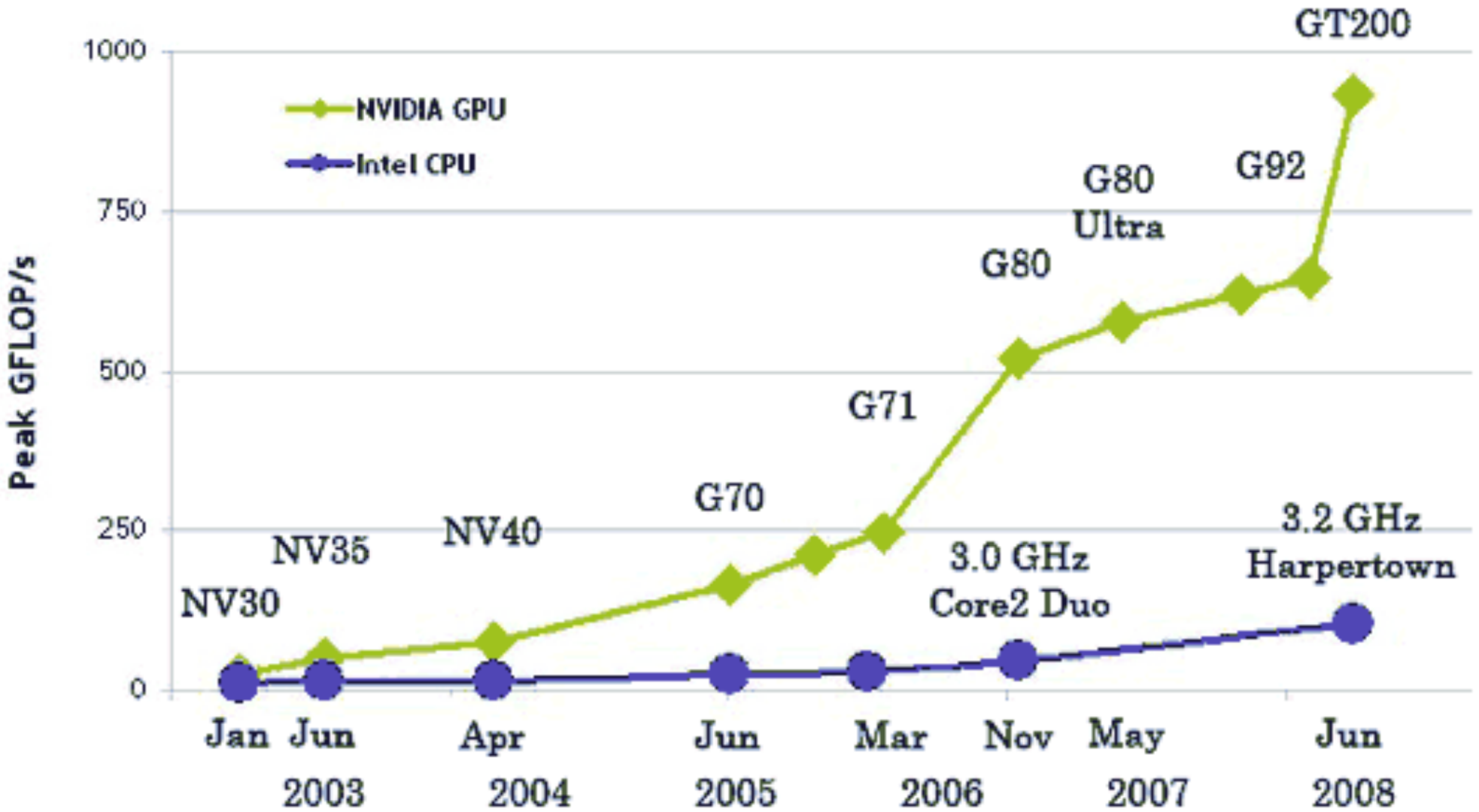


Figure 7: AMD Radeon™ HD 7970




100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Alpha Blend
Pixel Shade
Pixel Setup
Depth Test
Raster-ization
Vertex Shade
Pre-Vertex


GT200
1000
NVIDIA GPU
Intel CPU
750
G80
Ultra
G92
G80
500
G71
Peak GFLOP/s
G70
250
3.2 GHz
Harpertown
NV35
NV40
3.0 GHz
Core2 Duo
NV30
0
Jan
Jun
Apr
Jun
Mar
Nov
May
Jun
2003
2004
2005
2006
2007
2008

What next?