



Jon Peddie

# The History of the GPU - Steps to Invention

 Springer

# The History of the GPU - Steps to Invention

Jon Peddie

# The History of the GPU - Steps to Invention

 Springer

Jon Peddie  
Jon Peddie Research  
Tiburon, CA, USA

ISBN 978-3-031-10967-6                      ISBN 978-3-031-10968-3 (eBook)  
<https://doi.org/10.1007/978-3-031-10968-3>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Foreword

History often elicits strong responses whether it is studied in school, the subject of documentary films and books, or passed orally from generation to generation. No matter the source, no history can cover every event for any one person. My own memory demonstrates that daily.

I believe that history is an essential subject. Understanding what happened in the past gives insight into what worked and (perhaps more importantly) what didn't work and why. In addition, history provides context for current events. We learn from history in important ways.

Computing itself is a relatively new field. Many science and engineering fields are significantly older and their history has been documented extensively. There are substantive debates about what counts as the first digital computer. Suffice it to say that digital computers are not much more than 100 years old.

Computer graphics is an even newer field. It integrates disparate display technologies, digital and analog computers, and a human's innate capability to see pictures on a flat screen. Verne Hudson from Boeing-Wichita coined the term circa 1960. His collaborator, Bill Fetter, popularized it.

Jon's book complements a spate of recent publications devoted to the history of different aspects of computer graphics. Books by Peddie, Masson, and Carlson describe the field in general. Smith traces the evolution of the pixel. Llach looks at graphics in building and architecture, Weisberg the history of CAD, and Gaboury the influence of the University of Utah. This is just a sampling.

What I find interesting about the authors is that many are intimately involved with the field rather than historians per se. A number of them are pioneers or students of pioneers who have first-hand knowledge of the history they are documenting. These authors write with both authority and immediacy.

This book provides a broad view of the graphics processing unit. Jon has been involved with special purpose graphics processing technology since day one. He does an excellent job documenting processors dedicated to generating better images faster. Like any history, it's not complete. The book does provide a coherent, well-organized view of the evolution of a valuable technology. Jon emphasizes how GPUs evolved from custom processors devoted to picture generation to general-purpose parallel

processors. It provides context that helps the reader better understand how GPUs fit into the computer graphics world.

I was totally unaware of Hudson and Fetter and the existence of computers and computer graphics until the late 1960s. I didn't enter high school until 1962. My curriculum included Latin, Greek, and little science. Therefore, I could barely spell "computer." Ironically, I retired from Boeing as a Senior Technical Fellow in visualization and interactive techniques after a 35-year career.

The computer graphics bug bit me as a Johns Hopkins undergrad in 1969. Bill Huggins, who had spent his sabbatical learning computer animation with Bell Labs pioneers, recruited me to make computer-animated educational films. The process was arduous. It involved punched cards, line printer keyframes, a microfilm recorder (located in Brooklyn NY), an assembly language animation "language," and an IBM 7094 mainframe. There were no interactive devices for animators/programmers, no color output, no shaded images, and no sound. Just white lines on a black background. And I loved it!

My early career let me create more animated films and learn about interactive graphics at Battelle-Columbus Labs. I became aware that a digital computer can display one frame at a time whether the frame is part of a projected film or displayed on a graphics screen. The human visual system does the rest and gives a person the illusion of continuous motion as long as each image is shown quickly enough.

For the film, a projector shows frames fast enough (24–30 Hz) to make the motion seem continuous. Images on interactive device screens must be redrawn at the same rate or faster. Current interactive devices established a redraw rate at 60+ Hz. The requirement to draw new frames interactively ultimately led to the work with GPUs. A film may take compute-centuries to produce enough frames for a full-length animated film. Projectors are responsible for showing the frames fast enough.

GPUs help reduce compute-centuries for a film to something more reasonable by improving overall throughput. Interactivity pushes compute performance even harder. In today's interactive graphics world, GPUs must compute a completely new frame fast enough to create the illusion of continuous motion. Put another way, the image generation compute task, the task GPUs perform, must determine the color of each pixel on each frame fast enough to convince the human visual system that image transformations (either 2D or 3D) are continuous.

My work at Boeing emphasized acceptable interactive performance. I was able to work at a Boeing scale (interactively working with the complete digital design of a commercial airplane like a 787, ~2 billion polygons) on a GPU-equipped PC to make end-users think the task was easy. I often measure success by making the difficulty of complicated behind-the-scenes tasks seem simple when in actual use.

I think Jon's discussion about GPU evolution to become a generalized parallel processor adds real value. It confirms my belief that the most successful and powerful technologies are those that can be generalized and applied to problems the original developers never foresaw. GPUs fit that profile.

Pay careful attention to the lessons learned from GPU evolution and generalization. Those lessons can be applied to the reader's own work. And understand how forthcoming generations of GPUs can be extended to provide even more value in the future.

Sammanish, WA, USA  
June 2022

D. J. Kasik

# Preface

This is the first book in the three-book series on the History of the GPU.

History books are challenging to write. Technical history books are incredibly challenging. Why? Because things don't happen in an orderly sequence. Although one might think that event A leads to event B, often A leads to D, and B leads to C, but C leads to G.

Because the integrated graphics processing unit (GPU) has been employed in so many systems (platforms) and evolved since 1996, how do you tell a 2D story in a linear presentation such as the book?

One possibility is to list everything chronologically. Another approach is to list things by platform. And yet another choice is to list items by company, or by applications.

I have chosen a combination of all three.

This first book in the series covers the developments that lead up to the integrated GPU, from the early 1960s to the late 1990s

The book has two main sections, the PC platform and other platforms. Other platforms include workstations and game machines.

Each chapter is designed to be read independently, hence there may be some redundancy. Hopefully, each one tells an interesting story.

In general, a company is discussed and introduced in the year of its formation. However, a company may be discussed in multiple time periods in multiple chapters depending on how significant their developments were and what impact they had on the industry.



History of the GPU		
Steps to Invention Book 1	Eras and Environment Book 2	New Developments Book 3
1. Preface	1. Preface	1. Preface
2. History of the GPU	2. Race to build the first GPU	2. Second Era of GPUs (2001-2006)
3. 1980-1990 Graphics Controllers on Other Platforms	3. GPU Functions	3. Third to Fifth Era of GPUs
4. 1980-1989 Graphics Controllers on PCs	4. Major Era of GPUs	4. Mobile GPUs
5. 1990-1995 Graphics Controllers on PCs	5. First Era of GPUs	5. Game Console GPUs
6. 1990-1999 Graphics Controllers on Other Platforms	6. GPU Environment-Hardware	6. Compute GPUs
7. 1996-1999 Graphics Controller on PCs	7. Application Program Interface (API)	7. Open GPUs
8. What is a GPU	8. GPU Environment-Software Extensions	8. Sixth Era of GPUs

The History of the GPU - Steps to Invention

I mark the GPU’s introduction as the first fully integrated single chip with hardware geometry processing capabilities—transform and lighting. Nvidia gets that honor on the PC by introducing their GeForce 256 based on the NV10 chip in October 1999. However, Silicon Graphics Inc. (SGI) introduced an integrated GPU in the Nintendo 64 in 1996, and ArtX developed an integrated GPU for the PC a month after Nvidia. As you will learn, Nvidia did not introduce the concept of a GPU, nor did they

develop the first hardware implementation of transform and lighting. But Nvidia was the first to bring all that together in a mass-produced single-chip device.

The evolution of the GPU did not stop with the inclusion of the transformation and lighting (T&L) engine because the first era of such GPUs had fixed-function T&L processors—that was all they could do and when they were not doing that they sat idle using power. The GPU kept evolving and has gone through six eras of evolution ending up today as a universal computing machine capable of almost anything.

However, to fully appreciate and hopefully understand what wonderful development the GPU has been, it is necessary to know where, why, and how it was developed. To do that I start the story with the early computers from the late 1950s and 1960s.

Now GPUs are ubiquitous.

## **What Is In and Not In These Books**

As a public speaker and former engineer, you can tell from the above diagram; I like block diagrams. I have attempted to illustrate all the innovative GPUs and some of their predecessors with block diagrams. In some cases, I could not find sufficient data to construct a diagram; in some cases, the best I could do was a system-level diagram where the GPU is just a block.

In these books, you won't find any formulas (no math), code examples, operating application examples, user interface illustrations, and hopefully no commercials or propaganda.

Notable quotes and long quotations are presented indented to identify them as important and separate from the text.

At the end is the glossary. Not every term used in the book is in the glossary as many of the explanations are in the body text.

There is also a list of acronyms. The tech industry loves acronyms, and they can save time in communicating; they can also be very confusing. The acronym lists the acronym and a brief description.

## **Significant Things**

One of my goals for these books was to identify those developments that I (and hopefully others) thought were inflection points and disruptive results—things that moved the industry and or changed its direction. I marked those milestones in bold italics.

The introduction of the GPU was just such a thing. It has profoundly and forever changed how computers work and are used.

I hope you find this and the following books interesting and informative. I have personally lived through almost all of it and have known most of the people

mentioned. Many are acquaintances, and many are friends. Many of the people mentioned have generously contributed to this book with fact-checking, storytelling, and encouragement. However, it's necessary to say that any mistakes or inaccuracies are all my own.

## The Author

### *A Lifetime of Chasing Pixels*

I have been working in computer graphics since the early 1960s, first as an engineer, then as an entrepreneur (I found four companies and ran three others), ending up in a failed attempt at retiring in 1982 as an industry consultant and advisor. Over the years, I watched, advised, counseled, and reported on developing companies and their technology. I saw the number of companies designing or building graphics controllers swell from a few to over forty-five. In addition, there have been over thirty companies designing or making graphics controllers for mobile devices.

I've written and contributed to several other books on computer graphics (seven under my name and six co-authored). I've lectured at several universities around the world, written uncountable articles, and acquired a few patents, all with a single, passionate thread—computer graphics and the creation of beautiful pictures that tell a story. This book is liberally sprinkled with images—block diagrams of the chips, photos of the chips, the boards they were put on, and the systems they were put in—and pictures of some of the people who invented and created these marvelous devices that impact and enhance our daily lives—many of them I am proud to say are good friends of mine.

I laid out the book in such a way (I hope) that you can open it up to any page and start to get the story. You can read it linearly; if you do, you'll probably find a new information and probably more than you ever wanted to know. My email address is in various parts of this book, and I try to answer every one, hopefully within 48 h. I'd love to hear comments, your stories, and your suggestions.

The following is an alphabetical list of all the people (at least I hope it's all of them) who helped me with this project. A couple of them have passed away, sorry to say. Hopefully, this book will help keep the memory of them and their contributions alive.

Thanks for reading

Jon Peddie—*Chasing pixels, and finding gems*

## Acknowledgments and Contributors

The following people helped me with editing, interviews, data, photos, and most of all encouragement. I literally and figuratively could not have done this without them.

Anand Patel—Arm  
Andrew Wolfe—S3  
Ashraf Eassa—Nvidia  
Atif Zafar—Pixilica  
Borger Ljosland—Falanx  
Brian Kelleher—DEC, and finally Nvidia  
Bryan Del Rizzo—3dfx & Nvidia  
Carrell Killebrew—TI/ATI/AMD  
Chris Malachowsky—Nvidia  
Curtis Priem—Nvidia  
Dado Banatao—S3  
Dan Vivoli—Nvidia  
Dan Wood—Matrox, Intel  
Daniel Taranovsky—ATI  
Dave Erskine—ATI & AMD  
Dave Kasik—Boeing  
Dave Orton—SGI, ArtX, ATI & AMD  
David Harold—Imagination Technologies  
Edvaed Sergard—Falanx  
Emily Drake—Siggraph  
Eric Demers—AMD/Qualcomm  
Frank Paniagua—Video Logic  
Gary Tarolli—3dfx  
George Sidiropoulos—Think Silicon  
Gerry Stanley—Real3D  
Henry C. Lin—Nvidia  
Henry Chow—Yamaha & Giga Pixel

Henry Fuchs—UNC  
Henry Quan—ATI  
Hossain Yassaie—Imagination Technologies  
Iakovos Istamoulis—Think Silicon  
Ian Hutchinson—Arm  
Jay Eisenlohr—Rendition  
Jay Torberg—Microsoft  
Jeff Bush—Nyuzi  
Jeff Fischer—Weitek & Nvidia  
Jem Davis—Arm  
Jensen Huang—Nvidia  
Jim Pappas—Intel  
Joe Curley—Tseng/Intel  
John Poulton—UNC & Nvidia  
Jonah Alben—Nvidia  
Karl Gutttag—TI  
Karthikeyan (Karu) Sankaralingam—University of Wisconsin-Madison  
Kathleen Maher—JPA & JPR  
Ken Potashner—S3 & SonicBlue  
Kristen Ray—Arm  
Lee Hirsch—Nvidia  
Luke Kenneth Casson Leighton—Libre-GPU  
Mark Kilgard—Nvidia (Iris GL)  
Mary Whitton—Iknoas  
Megan Zea—PCI SIG  
Melissa Scuse—Arm  
Mike Diehl—HP  
Mike Mantor—AND  
Mikko Alho—Siru  
Mikko Nurmi—Bitboys

Neal Leavitt—Editing

Neil Trevett—3Dlabs & Khronos

Nick England—Iknoas

Pedro Duarte—Universities of Coimbra

Peter McGuinness—SGS Thompson

Peter L. Segal—AT & T

Petri Norlund—Bitboys

Phil Roges—ATI

Richard Huddy—ATI

Richard Selvaggi—Tseng Labs

Rick Bergman—ATI/AMD

Robert Dow—JPR

Ross Smith—3dfx

Ruchika Saini—editing

Sasa Marinkovic—ATI & AMD

Simon Fenny—Video Logic & Imagination Technologies

Stefan Demetrescu—Stanford

Stephen Morein—Stellar

Steve Brightfield—SiliconArts

Steve Edelson—Edson Labs

Tatsuo Yamamoto—Sega/DMP

Tim Leland—Qualcomm

Timothy Miller—Traversal Technology

Tom Forsyth—3Dlabs

Tony Tamasi—3dfx & Nvidia

Trevor Wing—Video Logic

# Contents

<b>1</b>	<b>Introduction</b>	1
1.1	Introduction	1
1.2	First Computer Graphics System (1949)	4
1.3	The Graphics Processor Unit (1999)	12
1.3.1	The Evolution of Graphics Controllers to GPUs	14
1.4	Performance (2000–2026)	16
1.5	The GPU’s Changing Role	17
1.6	The GPU’s Application	19
1.6.1	AI and Machine Learning	19
1.6.2	Accelerated Computing and Supercomputers	20
1.6.3	Content Creation	20
1.6.4	Gaming	20
1.6.5	Molecular Modeling	21
1.6.6	Video and Photo Editing	21
1.6.7	Vehicle Navigation and Robots	22
1.6.8	Crypto Mining	22
1.6.9	Summary	22
1.7	The Many Roles of the GPU Require Additional Names	22
1.8	Types of GPUs	26
1.9	Conclusion	28
	References	28
<b>2</b>	<b>1980–1989, Graphics Controllers on Other Platforms</b>	31
2.1	Ikonas Graphics Systems (1978–1982)	34
2.2	Pixel Planes—The Foundation of the GPU (1980–2000)	39
2.2.1	HP Acquires Division (1996)	48
2.3	Processor per Polygon—The Demetrescu Caltech Architecture (1980)	60
2.4	The Geometry Engine (1981)	63
2.5	Matrox SM640 with Geometry Engine (1987)	67

- 2.6 SGI’s Personal Integrated Raster Imaging System (IRIS) Workstation (1988) . . . . . 67
- 2.7 SGI’s IrisVision AIB (1988) . . . . . 69
- 2.8 NEC’s  $\mu$ PD7220 Pioneering Graphics Display Controller (1982) . . . . . 69
- 2.9 Hitachi ACRTC HD63484 (1984) . . . . . 73
- 2.10 Truevision (1984–1987) . . . . . 77
- 2.11 TI 34010 (1986) . . . . . 78
  - 2.11.1 TI Epilogue . . . . . 85
- 2.12 MAGIC—Multiple Application Graphics Integrated Circuit (1987) . . . . . 85
- 2.13 Raster Technologies Vertex Processor (1987) . . . . . 87
- 2.14 Amiga (1988) . . . . . 88
- 2.15 Sun’s GX Graphics Accelerator Board (1989) . . . . . 91
  - 2.15.1 Summary . . . . . 94
- 2.16 Conclusion . . . . . 94
- References . . . . . 95
- 3 1980–1989, Graphics Controllers on PCs . . . . . 99**
  - 3.1 1980–1989, Graphics Controllers on the PC Platform . . . . . 99
  - 3.2 CRT Control (1975–1987) . . . . . 102
    - 3.2.1 The Video Output—LUT-DAC (~1981–1987) . . . . . 102
    - 3.2.2 Brooktree (1983–1996) . . . . . 103
    - 3.2.3 Edsun Labs (1989–1991) . . . . . 104
    - 3.2.4 Summary of Video Output . . . . . 108
  - 3.3 IBM Graphics History (1981–1990) . . . . . 108
    - 3.3.1 IBM CGA (1981) . . . . . 109
    - 3.3.2 IBM EGA (1984) . . . . . 109
    - 3.3.3 EGA Begets VGA to XGA . . . . . 111
    - 3.3.4 The IBM Professional Graphics Controller—PGC (1984) . . . . . 112
    - 3.3.5 The IBM 8514/A (1987) . . . . . 114
    - 3.3.6 IBM VGA (1987–1991) . . . . . 116
    - 3.3.7 Those Clones . . . . . 120
    - 3.3.8 IBM Summary . . . . . 121
  - 3.4 The Market Expands (1986–1987) . . . . . 121
  - 3.5 Intel’s Pre-GPU History (1983–2003) . . . . . 122
    - 3.5.1 82720 (1983) . . . . . 122
    - 3.5.2 82786 (1986) . . . . . 122
    - 3.5.3 i860 (1989) . . . . . 124
    - 3.5.4 i740 (1998) . . . . . 125
    - 3.5.5 i810 (1999) . . . . . 126
    - 3.5.6 Extreme Graphics (2001) . . . . . 128
    - 3.5.7 Intel Summary . . . . . 129
  - 3.6 Tseng Labs (1983–1997) . . . . . 129



- 3.6.1 Winning Awards Was not Enough ..... 135
- 3.6.2 It Could Have Been the First GPU ..... 136
- 3.6.3 The End ..... 137
- 3.7 SGI’s IrisVision (1988–1994) ..... 138
  - 3.7.1 The Legacy of IrisVision—Pellucid, Media Vision,  
and 3dfx (1991–1994) ..... 141
  - 3.7.2 Media Vision (1990–1994) ..... 142
  - 3.7.3 Benchmarking ..... 143
- 3.8 Conclusion ..... 144
- References ..... 144
- 4 1980–1995 the Progenitors: Graphics Controller on PCs ..... 147**
  - 4.1 1990–1995, Graphics Controllers on PCs ..... 147
    - 4.1.1 IBM XGA (1990) ..... 147
    - 4.1.2 Summary 1990 to 1995 ..... 152
  - 4.2 The IGC to IGP (1991) ..... 152
    - 4.2.1 The First Workstation IGC ..... 153
    - 4.2.2 The First PC IGC ..... 153
  - 4.3 Bitboys (1991–1999) ..... 154
    - 4.3.1 Pyramid3D 25202 ..... 157
    - 4.3.2 Pyramid3D 25201 ..... 158
    - 4.3.3 The Eight *P*’s ..... 160
    - 4.3.4 Summary ..... 162
  - 4.4 Artist Graphics (1979–2098) ..... 162
    - 4.4.1 Artist Graphics Shows 3GA Graphics Accelerator ..... 163
    - 4.4.2 Summary ..... 165
  - 4.5 Number Nine Imagine 128 (1992–1999) ..... 166
    - 4.5.1 Summary ..... 169
  - 4.6 Rendition (1992–1998) ..... 169
    - 4.6.1 Summary ..... 176
  - 4.7 Stellar—RSSI (1993–2000) ..... 177
    - 4.7.1 Reality Simulations Systems PixelSquirt ..... 179
    - 4.7.2 Stellar is Born (1997) ..... 181
    - 4.7.3 VelaTX (1998) ..... 182
    - 4.7.4 Broadcom Acquires Stellar (2000) ..... 183
    - 4.7.5 Summary ..... 184
  - 4.8 Matrox Millennium (1994–2014) ..... 184
    - 4.8.1 Summary ..... 188
  - 4.9 VideoLogic/Imagination Technologies Tiling (1994–) ..... 188
    - 4.9.1 NEC-Imagination Technologies PCX (1994–1999) ..... 193
    - 4.9.2 Summary ..... 198
  - 4.10 Conclusion ..... 200
  - References ..... 200

- 5 1990 to 1999 Graphics Controllers on Other Platform . . . . . 203**
  - 5.1 Workstations . . . . . 203
    - 5.1.1 Workstation Graphics . . . . . 204
    - 5.1.2 HP Artist (1993) . . . . . 205
    - 5.1.3 Silicon Reality (1994–1998) . . . . . 209
    - 5.1.4 The Saga of Evans & Sutherland’s Pre-GPU Effort  
(1995–2001) . . . . . 214
    - 5.1.5 3Dlabs Permedia (1997) . . . . . 222
    - 5.1.6 Intergraph Wildcat (1998–2000) . . . . . 230
  - 5.2 Game Consoles . . . . . 234
    - 5.2.1 Sega . . . . . 235
    - 5.2.2 Sega Genesis (1988) . . . . . 235
    - 5.2.3 Sony PlayStation (1994) . . . . . 235
    - 5.2.4 Atari Jaguar (1993) . . . . . 240
    - 5.2.5 Nintendo 64 (1996)—The First T&L in a Console . . . . . 243
    - 5.2.6 ArtX and the Nintendo GameCube (1998) . . . . . 250
    - 5.2.7 NEC Electronics’ PowerVR (1996) . . . . . 253
  - 5.3 Conclusion . . . . . 261
  - References . . . . . 262
  
- 6 1996–1999, Graphics Controllers on PCs . . . . . 265**
  - 6.1 The ATI 3D Rage (1995) . . . . . 265
    - 6.1.1 Approaching the GPU . . . . . 269
    - 6.1.2 The Saga of ATI (1985–2006) . . . . . 273
  - 6.2 Nvidia’s Quadratic Processor, the NV1 (1993–) . . . . . 275
    - 6.2.1 Nvidia Epilogue . . . . . 282
  - 6.3 3dfx Voodoo (1994–2000) . . . . . 283
    - 6.3.1 SLI Was Not a New Concept . . . . . 290
    - 6.3.2 The SST-1 . . . . . 290
  - 6.4 Yamaha YGV612 RPA (1995–1996) . . . . . 300
  - 6.5 Real3D (1995–1999) . . . . . 303
    - 6.5.1 A Stand-Alone Company . . . . . 306
    - 6.5.2 Real3D and Silicon Graphics Settle Out of Court . . . . . 307
    - 6.5.3 Intel Acquires Real3D (October 25, 1999) . . . . . 309
    - 6.5.4 3dfx and Intel Patent Cross-License Agreement . . . . . 311
  - 6.6 Microsoft Talisman—The Chip That Never Was (1996) . . . . . 311
  - 6.7 Nvidia Riva 128 (1996) . . . . . 319
  - 6.8 S3 Virge 86C385 (1996) . . . . . 322
    - 6.8.1 S3 Epilogue . . . . . 328
  - 6.9 Conclusion . . . . . 329
  - References . . . . . 330

- 7 What is a GPU?** ..... 333
  - 7.1 What is a GPU? ..... 333
  - 7.2 The GPU ..... 335
    - 7.2.1 Vendors ..... 336
    - 7.2.2 Shaders, Processors, Units, and Cores ..... 337
    - 7.2.3 Getting to a Model ..... 338
  - 7.3 The Six Eras of GPUs ..... 339
    - 7.3.1 Pre-GPU Graphics Controllers (1960–1998) ..... 340
    - 7.3.2 First-Era GPUs (1999–2000) Fixed Function ..... 341
    - 7.3.3 Second-Era GPUs (2000–2006) Programmable Shaders ..... 341
    - 7.3.4 Third-Era GPUs (2006–2009) Unified Shaders ..... 342
    - 7.3.5 Fourth-Era GPUs (2009–2015) Compute Shaders ..... 342
    - 7.3.6 Fifth-Era GPUs (2015–2020) Ray Tracing and AI ..... 342
    - 7.3.7 Sixth-Era GPUs (2020–) Mesh Shaders ..... 343
    - 7.3.8 The Range of the GPU and This Book ..... 343
  - 7.4 Conclusion ..... 344
  - 7.5 Epilog ..... 344
  - References ..... 345
  
- Appendix A: Acronyms** ..... 347
- Appendix B: Definitions** ..... 353
- Index** ..... 393

# List of Figures

- Fig. 1.1 A raster graphics display consists of quantized elements known as pixels ..... 3
- Fig. 1.2 The small-scale experimental machine (SSEM), called baby, was built at the University of Manchester in June 1948 ... 4
- Fig. 1.3 Baby’s dot-matrix display ..... 5
- Fig. 1.4 Whirlwind—the first interactive digital computer. Stephen Dodd (sitting), Jay Forrester, Robert Everett, and Ramona Ferenz at the Whirlwind I, test control display in the Barta Building, 1950 (Courtesy of The MITRE Corporation) ..... 6
- Fig. 1.5 Using a light gun on a SAGE air defense screen to pick a target aircraft (Courtesy of IBM) ..... 7
- Fig. 1.6. 3D perspective drawing created by William Fetter at Boing (Courtesy of McGraw-Hill) ..... 7
- Fig. 1.7 Ivan Sutherland demonstrating Sketchpad (Courtesy of Wikipedia) ..... 8
- Fig. 1.8 The first stand-alone workstation, IDI’s IDDIOM with Calligraphic screen and light-pen (Courtesy of IEEE) [16] ..... 9
- Fig. 1.9 An engineer using a light pen on a Control Data 274 Digigraphics vector display terminal, circa 1965 (Courtesy of the Charles Babbage Institute Archives, University of Minnesota Libraries) ..... 9
- Fig. 1.10 A Tektronix graphics terminal system board (Courtesy of Legalizeadulthood) ..... 11
- Fig. 1.11 Wolfenstein 3D was the first PC-based 3D first-person shooter (Courtesy of Software & Apogee Software 1992 id) .... 12
- Fig. 1.12 The evolutionary path of the GPU ..... 14
- Fig. 1.13 The entire image above was calculated; it is not a photograph or texture map (Courtesy of Epic Games, Nanite demo) ..... 15
- Fig. 1.14 Performance of popular platforms over time ..... 16

Fig. 1.15 Nvidia’s GeForce 256, the first single-chip GPU (Courtesy of Konstantin Lanzet, Wikipedia) . . . . . 17

Fig. 1.16 GPUs have become ubiquitous and accelerated science, resulting in new products, enhanced vehicle safety, and many other applications . . . . . 18

Fig. 1.17 Taxonomy of names . . . . . 25

Fig. 1.18 GPUs are found in many types of systems and have different prefixes . . . . . 26

Fig. 1.19 The problems with segmentations and names . . . . . 27

Fig. 2.1 Monochrome 2D line drawing done with a NEC 7220, on a PC ruining AutoCAD circa in the early 1980s (Historic American Buildings Survey: Library of Congress) . . . . . 32

Fig. 2.2 Imaginary worlds in color 3D games run on graphics chips circa 1990 (Courtesy of Wikipedia) . . . . . 32

Fig. 2.3 Historical view of the generic organization of 2D/3D raster systems . . . . . 34

Fig. 2.4 Ikonas graphics system block diagram . . . . . 36

Fig. 2.5 Nick England and Mary Whitton recreating a 1980 Ikonas booth at Siggraph’s 25th anniversary in 1998 (Courtesy of England) . . . . . 37

Fig. 2.6 The Ikonas system (Courtesy of England) . . . . . 37

Fig. 2.7 Tim Van Hook’s ray tracing code rendering bi-cubic B-spline and polygonal surfaces (Courtesy of Nick England) . . . . . 38

Fig. 2.8 Tim Van Hook, Fellow, ATI Technologies 2001 (Courtesy of Nick England) . . . . . 39

Fig. 2.9 Fuchs’s background involved modeling chromosomes on graphics systems. He constructed 3D models from laser scans of people and objects before coming to UNC in 1978 (Courtesy of UNC Endeavors) . . . . . 40

Fig. 2.10 Pixel planes overview block diagram—the template for the GPU . . . . . 42

Fig. 2.11 Pixel Planes’ functional organization, the ring was the key to communications . . . . . 44

Fig. 2.12 Pixel planes memory chip block diagram . . . . . 45

Fig. 2.13 Professor Henry Fuchs manipulates joysticks on the pixel planes system while an associate holds a memory board in front (Courtesy of Department of Computer Science, University of North Carolina) . . . . . 46

Fig. 2.14 PixelFlow prototype system block diagram . . . . . 47

Fig. 2.15 PixelFlow system organization . . . . . 49

Fig. 2.16 PixelFlow board . . . . . 49

Fig. 2.17 EMC memory chip in PixelFlow system was segmented . . . . . 51

Fig. 2.18 The long 30-year trail (and tale) of pixel planes . . . . . 59

Fig. 2.19 Dr. Professor Henry Fuchs, the father of the GPU  
(Courtesy of Department of Computer Science, University  
of North Carolina) ..... 59

Fig. 2.20 Stefan Demetrescu (Courtesy of Lasergraphics) ..... 60

Fig. 2.21 Cohen and Demetrescu’s pipelined polygon architecture ..... 61

Fig. 2.22 Competing surface processors acted much like a dataflow  
machine ..... 62

Fig. 2.23 James Clark (Courtesy of IEEE Computer) ..... 64

Fig. 2.24 Dot product ..... 64

Fig. 2.25 Three basic operations performed by a graphics system:  
transformation, clipping, and scaling ..... 65

Fig. 2.26 A block diagram of the Geometry Engine corresponding  
to the photo in Fig. 2.27 ..... 66

Fig. 2.27 Photograph of the Geometry Engine (Courtesy of ACM  
0-89791-076-1/82/007/0127) ..... 66

Fig. 2.28 Matrox SM 640 was the first 3D PC AIB and used SGI’s  
Geometry Engine (Courtesy of Matrox) ..... 67

Fig. 2.29 SGI’s IRIS 2000 graphics workstation, circa 1985  
(Courtesy of wiki.preterhuman.net) ..... 68

Fig. 2.30 NEC’s  $\mu$ PD7220 was the first integrated graphics  
controller chip ..... 70

Fig. 2.31 Layout of the  $\mu$ PD7220—notice the (dark) RAM area  
(Courtesy of Nikkei) ..... 71

Fig. 2.32 Block diagram of NEC’s  $\mu$ PD7220 GDC ..... 72

Fig. 2.33 Hitachi HD63484 ACRTC, more functionality and larger  
than the 7220 ..... 74

Fig. 2.34 An ELSA workstation add-in board using the Hitachi  
HD63484, the top row of chips are memory (Courtesy  
of VGA Museum) ..... 74

Fig. 2.35 A Force Computer VME SYS68k/AGC-1A add-in board  
based on the Hitachi HD63484 chip (Courtesy of Force  
Computers) [52] ..... 76

Fig. 2.36 Block diagram of the Hitachi HD63484 graphics controller .... 77

Fig. 2.37 Karl Gutttag (Courtesy of Gutttag) ..... 79

Fig. 2.38 The TMS34010’s system block diagram ..... 80

Fig. 2.39 The TMS34010’s internal architecture block diagram ..... 81

Fig. 2.40 A Spea TI TMS34010-based AIB with a memory  
at the top, a VGA (clone) chip onboard and a TI LUT-DAC  
(Courtesy of Konstantin Lanzet Wikipedia) ..... 82

Fig. 2.41 A photograph of the Texas Instruments’ TMS3020  
Graphics System Processor die (Courtesy of Pauli  
Rautakorpi, Wikipedia) ..... 83

Fig. 2.42 The University of Sussex’s MAGIC used in a system ..... 86

Fig. 2.43 Jay Torborg (Courtesy of Velotech) ..... 87

Fig. 2.44 A typical graphics command processing data flow ..... 88

Fig. 2.45	The Commodore Amiga block diagram	89
Fig. 2.46	A picture in the HAM mode, showing all 4,096 colors at once on-screen. Such an image was displayed on an Amiga 1000 in 1985! (Courtesy of The Amiga Museum)	90
Fig. 2.47	Sun Microsystem's GX graphics accelerator AIB functions	92
Fig. 2.48	Sun Microsystem's GX AIB (Courtesy of Curtis Priem)	93
Fig. 3.1	The IBM PC circa 1981 (Courtesy of Wikipedia, Rama & Musée Bolo)	100
Fig. 3.2	Brooktree LUT-DAC	103
Fig. 3.3	Brooktree LUT-DAC chip (Courtesy of Thomas Schanz Wikipedia)	104
Fig. 3.4	Steve Edelson, Edsun Laboratories (Courtesy of Edelson)	105
Fig. 3.5	A breadboard with wire-wrap pins (Courtesy of Russ Shumaker)	106
Fig. 3.6	Edsun's triangle demo: The bitmap was animated and rainbow-rotated around the edges (Courtesy of Steve Edelson)	106
Fig. 3.7	Analog devices/Edsun Labs CEG/DAC	107
Fig. 3.8	The CEG/DAC accepted processor and pixel data	107
Fig. 3.9	IBM's CGA add-in board (Courtesy of Wikipedia)	109
Fig. 3.10	IBM EGA add-in board. Notice the similarity to the CGA in form factor and layout (Courtesy of Vlask)	110
Fig. 3.11	The integrated EGA controller reduced the size of the AIBs of the era (Courtesy of VGA Museum)	110
Fig. 3.12	The 4-bit RGBI palette added an intensity bit	111
Fig. 3.13	Three-board-set of IBM's Professional Graphics Controller (PGC) (Courtesy of John Elliot Vintage PCs)	113
Fig. 3.14	Block diagram of IBM PGC with microprocessor and graphics emulation	113
Fig. 3.15	IBM's 8514/A AIB (lower) and memory board (above) formed a sandwich (Courtesy of os2museum.com)	114
Fig. 3.16	Block diagram of IBM's 8514/A in system	115
Fig. 3.17	The organization of a VGA/8514/A system	116
Fig. 3.18	IBM's highly integrated motherboard-based VGA chip (Courtesy of Wikipedia)	117
Fig. 3.19	A VGA board with EISA tab (top) and ISA tab (bottom); note VGA connector on each end of the board (Courtesy of ELSA/Wikipedia)	118
Fig. 3.20	IBM VGA block diagram	119
Fig. 3.21	The ubiquitous 15-pin VGA connector	120
Fig. 3.22	History of VLSI graphics chips	121
Fig. 3.23	The Intel SBX275 video graphics controller with 82,720 chip (Courtesy of Multibus International)	122

Fig. 3.24	Intel 82,786 die shot (Courtesy of <a href="https://Commons.wikimedia.org">https://Commons.wikimedia.org</a> )	123
Fig. 3.25	Intel i860 microprocessor (Courtesy of Wikipedia)	124
Fig. 3.26	Intel i740 prototype AIB with an AGP connector (Courtesy of <a href="http://www.SSSTjy.com">www.SSSTjy.com</a> )	126
Fig. 3.27	Intel i810 IGC	127
Fig. 3.28	Intel 810 chipset (Courtesy of Wikipedia)	128
Fig. 3.29	Intel's i845 Northbridge chipset (left) was surprisingly small (Courtesy of Wikipedia)	128
Fig. 3.30	Tseng Labs' ET4000AX. (Courtesy of Eep386: Wikipedia)	130
Fig. 3.31	Tseng Labs' ET4000 block diagram	131
Fig. 3.32	STB ET6000 AIB with 2 MB frame buffer and pad for additional MDRAM (Courtesy of Joe Curley)	132
Fig. 3.33	STB ET6000 block diagram	135
Fig. 3.34	SGI's IrisVision block diagram	138
Fig. 3.35	SGI's MCA-based IrisVision board-set interconnections (Courtesy of SGI)	139
Fig. 3.36	SGI's IrisVision AIB circa 1999 (Courtesy of eBay)	140
Fig. 3.37	Don Strimbu's Nozzle was a 2D drawing benchmark used for many years (Courtesy of CAD Nauseam)	144
Fig. 4.1	IBM XGA AIB (Courtesy of CC BY-SA 3.0, commons. Wikimedia)	148
Fig. 4.2	XGA block diagram, the coprocessor was the graphics engine	149
Fig. 4.3	IBM XGA-2 AIB (Courtesy of OS2 Museum)	151
Fig. 4.4	Integrated graphics controller circa 1992 (IGC)	153
Fig. 4.5	Bitboys' Pyramid3D AIB (Courtesy of Petri Nordlund)	155
Fig. 4.6	Bitboys 3D Graphics pipeline processing hierarchy	156
Fig. 4.7	Pyramid3D system architecture	156
Fig. 4.8	Bitboy's Pyramid3D architecture	159
Fig. 4.9	Symmetric multiprocessing with multiple Pyramid3D chips	161
Fig. 4.10	Artist Graphics's 3GA controller	164
Fig. 4.11	Number Nine's Imagine 128 PCI AIB circa 1995 (Courtesy of Wikipedia)	167
Fig. 4.12	Rendition team pose in front of their first official office in Mountain View California, 1994 (Courtesy of Jay Eisenlohr)	170
Fig. 4.13	Mike Boich (Courtesy of Mike Boice Wikipedia)	171
Fig. 4.14	Verité 2200 block diagram (outlined items new integrated components)	174
Fig. 4.15	Jay Eisenlohr (Courtesy of Engineering Oregon State)	177
Fig. 4.16	Pixel Squirt 3D core	180
Fig. 4.17	Matrox's first graphics AIB, the ALT-256, was designed in 1978 for early microcomputers (Courtesy of Matrox)	185



Fig. 4.18 Matrox Millennium ISA, circa 1997 (Courtesy of Gona.eu BY-SA 3.0 Wikipedia) . . . . . 186

Fig. 4.19 Sir Hossein Yassaie (Courtesy of The Times) . . . . . 189

Fig. 4.20 Tile-based deferred rendering (TBDR) pipeline . . . . . 190

Fig. 4.21 Martin Ashton (Courtesy of Ashton) . . . . . 191

Fig. 4.22 Simon Fenney (Courtesy of Fenney) . . . . . 192

Fig. 4.23 Prototype TBDR FPGA AIB (Courtesy of Simon Fenney) . . . . . 193

Fig. 4.24 VideoLogic’s Apocalypse 5D. (Courtesy of Fabian Günther-Borstel) . . . . . 194

Fig. 4.25 Planet of death (Courtesy of Ubisoft) . . . . . 196

Fig. 4.26 The Series2 PMX1, a prototype of what later became the Neon 250 (Courtesy of Imagination Technologies) . . . . . 199

Fig. 5.1 HP’s balanced compression/decompression with CPU and HP Artist chip . . . . . 206

Fig. 5.2 HP’s Artist chip block diagram . . . . . 207

Fig. 5.3 Signature in an integrated circuit chip (Courtesy of Florida State University’s Silicon Zoo project) . . . . . 209

Fig. 5.4 Silicon Reality’s TAZ Core function block diagram . . . . . 211

Fig. 5.5 Silicon Reality’s Tantrum block diagram . . . . . 212

Fig. 5.6 Silicon Reality’s Tantrum product functional block diagram . . . . . 213

Fig. 5.7 E&S Realimage DirectBurst graphics block diagram . . . . . 216

Fig. 5.8 Evans & Sutherland’s RealImage 2000 block diagram . . . . . 218

Fig. 5.9 Interleaved relationships in the late 1990s workstation market . . . . . 219

Fig. 5.10 3Dlabs’ Glint block diagram . . . . . 223

Fig. 5.11 3Dlabs’ Glint 300SX core architecture had many stages . . . . . 224

Fig. 5.12 The 3Dlabs’ Glint MX processor relied on an external geometry processor . . . . . 224

Fig. 5.13 3Dlabs’ road map in the mid-90s (Courtesy of 3Dlabs) . . . . . 226

Fig. 5.14 3Dlabs’ Permedia consumer-level 3D controller . . . . . 228

Fig. 5.15 The history of 3Dlabs . . . . . 230

Fig. 5.16 Elsa’s GLoria-XL 40-MB 3Dlabs-based workstation AIB (Courtesy of VGA Museum) . . . . . 231

Fig. 5.17 Intergraph’s Intense Wildcat (single pipeline) block diagram . . . . . 231

Fig. 5.18 The graphics from the Sega Genesis were not impressive, but in 1988 and for a standard definition TV, they rivaled arcade game machines (Courtesy of Wikipedia) . . . . . 236

Fig. 5.19 Ken Kutaragi (Courtesy of Wikipedia) . . . . . 236

Fig. 5.20 Sony’s 1991 PlayStation gamer developer demo with quasi-3D textures (Courtesy of [https://www.youtube.com/watch?v=rwNt\\_9GvpFI](https://www.youtube.com/watch?v=rwNt_9GvpFI)) . . . . . 238

Fig. 5.21 Comparison of affine and correct perspective texture mapping (Courtesy of Darkness3560 for Wikipedia) . . . . . 238

Fig. 5.22 PlayStation block diagram . . . . . 239

Fig. 5.23	Atari Jaguar block diagram	241
Fig. 5.24	Nintendo 64, the first console with a 3D accelerator (Courtesy of Wikipedia)	244
Fig. 5.25	Nintendo 64 motherboard, CPU, and reality coprocessor, with RDRAM below the processors (Courtesy of Nintendo)	244
Fig. 5.26	Nintendo 64 block diagram	245
Fig. 5.27	Nintendo 64 reality block diagram	246
Fig. 5.28	Nintendo 64 RCP block diagram	247
Fig. 5.29	Nintendo GameCube system board (Courtesy of Nintendo)	252
Fig. 5.30	Arcade implementation of NEC's VideoLogic ISP and TSP chips	254
Fig. 5.31	NEC/VL ISP block diagram	254
Fig. 5.32	NEC/VL TSP block diagram	255
Fig. 5.33	Trevor Wing (Courtesy of Register of Chinese Herbal Medicine)	256
Fig. 5.34	NEC/VL PCX1 block diagram	257
Fig. 5.35	Tatsuo Yamamoto (Courtesy of Yamamoto)	258
Fig. 5.36	Hideki Sato (Courtesy of Sega Retro)	259
Fig. 6.1	Kwok Yeun Ho, ATI founder and CEO (Courtesy of ATI)	266
Fig. 6.2	ATI's first graphics chip and board, the CW16800-A (Courtesy of TechPowerUp)	267
Fig. 6.3	ATI Vantage (left) and Ultra—notice the difference in the memory to the left of the graphics controller chips (Courtesy of VGA Museum)	269
Fig. 6.4	ATI's 3D Rage AIB (Courtesy of TechPowerUp)	270
Fig. 6.5	ATI 3D Rage II internal block diagram	272
Fig. 6.6	ATI's ImpacTV chip function block diagram	272
Fig. 6.7	History of ATI's acquisitions	275
Fig. 6.8	ATI's David Orton and AMD's Hector Ruiz officially announce the historic merger (Courtesy of AMD)	276
Fig. 6.9	Chris Malachowsky (Courtesy of Nvidia)	277
Fig. 6.10	Curtis Priem (Courtesy of Rensselaer)	277
Fig. 6.11	Diamond Multimedia's Edge 3D with SGS (Nvidia) chip (Courtesy of Wikipedia)	279
Fig. 6.12	Nvidia's NV1 block diagram	279
Fig. 6.13	Jensen Huang right after Sega delivered three arcade machines to Nvidia for testing and integration in 1995	281
Fig. 6.14	The founders of 3dfx: Scott Sellers, Gary Tarolli, and Ross Smith (Courtesy of Smith)	284
Fig. 6.15	The basic 3dfx graphics engine	288
Fig. 6.16	3dfx developed scan-line interleaving in 1995 (Courtesy of Martín Gamero Prieto)	289
Fig. 6.17	Block diagram of 3dfx's frame buffer interface chip	291
Fig. 6.18	The Voodoo1 from 3dfx, released in 1996 (Courtesy of Wikipedia)	292

Fig. 6.19	<i>Interstate '76</i> , released in 1997 by Activision, running on a Voodoo 1 (Courtesy of Wikipedia) . . . . .	293
Fig. 6.20	Comparison of 3dfx performance to VGA (Courtesy of 3dfx) . . . . .	296
Fig. 6.21	The last AIB from 3dfx, the Voodoo5 5500 (Courtesy of Wikipedia, Konstantin Lanzet) . . . . .	298
Fig. 6.22	The history of 3dfx . . . . .	298
Fig. 6.23	Yamaha YGY611 block diagram . . . . .	301
Fig. 6.24	Paradise's YGY612-based Tasmania AIB (Courtesy of Vogonwiki) . . . . .	303
Fig. 6.25	Gerald Stanley, Real3D's founder, and CEO (Courtesy of Stanley) . . . . .	304
Fig. 6.26	LMC's Real3D multichip Starfighter AIB . . . . .	306
Fig. 6.27	Memory bandwidth requirement (Mbytes/second) . . . . .	313
Fig. 6.28	Talisman system hardware partitioning . . . . .	313
Fig. 6.29	Talisman polygon object processor . . . . .	315
Fig. 6.30	Talisman image layer compositor . . . . .	316
Fig. 6.31	Nvidia's NV3 RIVA 128 media accelerator . . . . .	320
Fig. 6.32	Diamond Multimedia's RIVA 128, NV3-based Viper V330 4 MB gaming AIB circa 1995 (Courtesy of Mathías Tabó, Wikipedia) . . . . .	321
Fig. 6.33	Nvidia's RIVA 128 3D graphics engine . . . . .	321
Fig. 6.34	Dado Banatao, founder of S3 (Courtesy of Positively Filipino) . . . . .	323
Fig. 6.35	Terry Holdt would run S3 for ten years (Courtesy of Team 6502) . . . . .	324
Fig. 6.36	S3's ViRGE 86C385 3D graphics board (Courtesy of VGA Legacy) . . . . .	324
Fig. 6.37	The S3 ViRGE 86C385 block diagram . . . . .	325
Fig. 6.38	S3 Savage3D AIB (Courtesy of VGA Museum) . . . . .	327
Fig. 6.39	Ken Potashner, S3's final CEO (Courtesy of SONICblue) . . . . .	327
Fig. 6.40	S3's timeline . . . . .	329
Fig. 7.1	The basic GPU pipeline . . . . .	334
Fig. 7.2	The many elements of a modern GPU . . . . .	336
Fig. 7.3	A shaded triangle . . . . .	337
Fig. 7.4	Shaders, processors, clusters, and cores make up a GPU . . . . .	338
Fig. 7.5	The various SIMD clusters of the GPU suppliers . . . . .	338
Fig. 7.6	Raster or direct rendering uses CG tricks to approach realism and is not physically correct. . . . .	339
Fig. 7.7	The eras of GPU development . . . . .	341

# List of Tables

Table 1.1	Nvidia architectural names	23
Table 2.1	PixelFlow's custom chips characteristics	51
Table 3.1	Some of the VGA clone suppliers	120
Table 4.1	Specifications of the 3GA controller	165
Table 4.2	Matrox Mystique resolutions and refresh rates	187
Table 4.3	Imagination technologies PowerVR features	198
Table 5.1	Silicon Reality's TAZ core specifications	210
Table 5.2	Silicon Reality's TAZ 3D mode specifications	211
Table 5.3	Silicon Reality's Tantrum features	213
Table 5.4	Silicon Reality's 2D features	213
Table 5.5	Silicon Reality's Tantrum, range of display formats	214
Table 5.6	Game consoles before GPUs	235
Table 6.1	The family of versions of the ATI 3D Rage graphics controller	271
Table 6.2	Talisman characteristics and features	312

# Chapter 1

## Introduction



### 1.1 Introduction

Over the years, the computer's central processing unit (CPU) eventually incorporated every coprocessor developed to augment and add to its function except for one major processor, the GPU. A GPU is a specialized processor developed initially to accelerate graphics rendering and geometry transformations.

The CPU has even incorporated graphics processing. However, the CPU has not terminated the GPU's stand-alone value as it did with floating-point processors, digital signal processors (DSPs), video Codecs, and other accelerators. The GPU survives as a stand-alone coprocessor because the GPU scales almost infinitely—adding transistors to create thousands of processor cores. The only asymptote a GPU might face is inter-processor communications. Clustering groups of processors (shaders) overcomes that barrier. Coherent caches have also scaled well and address the GPU's inter-processor communications bottleneck. GPUs have been a significant beneficiary of Moore's law [3] (which postulates that the number of transistors on a chip doubles and cuts the price in half approximately every 1.5 to 2 years).

The GPU is a wonderful device and has made tremendous contributions to the computer's capabilities.

GPUs can process data simultaneously, which is known as parallel processing. As a result, GPUs are used in applications beyond gaming and simulation. Applications as far-ranging as artificial intelligence (AI), machine learning (ML), CAD, and compute-intensive tasks use GPUs. GPUs have been used as accelerators for photo and video editing and high-performance computers (HPC) and supercomputers.

GPUs were initially stand-alone discrete hardware units (dGPU). Later, in 2010, they were added to the CPU (iGPU) but still held their place as a stand-alone device. GPUs can contain specialized AI elements and multimedia accelerators for video and audio, and ray tracing accelerators. As the GPU advanced beyond its original role as a graphics processor and found use in pure computing applications not requiring a display, its additional capabilities have been called General-Purpose GPU (GPGPU), or GPU compute (cGPU). It is a false term because there is nothing general-purpose

about a GPU. It cannot run an operating system, manage disk drives and peripherals, or boot up a system, nor is there anything general-purpose about a parallel processor. Those are the jobs of the CPU. GPUs are specialized devices with specific and specialized parallel processing capabilities.

The terms *semiconductor*, *integrated circuit (IC)*, and *chip* will be used interchangeably in this book and should be considered synonyms.

What does a GPU do? It is almost everything needed to calculate in parallel from geometry processing to image processing to AI training and accelerated computing.

Computer graphics is about geometry; as Pixar cofounder Alvy Ray Smith says in his book, *A Biography of the Pixel*, “Computer graphics is geometry in, pixels out [4].” There’s much more to the GPU, however. Image processing is pixels in, pixels out, whereas AI training and compute-acceleration is data in, data out—and a GPU does all of that and more.

And David Kasik of Boeing says:

I consider computer graphics to be about creating, displaying, and modifying visual content to communicate to others. The sources are much broader than geometry: cameras, simulations, brain waves, sounds, etc. All forms of data can have a visual manifestation.

The integration of the rendering engine and the transformation and lighting (T&L) engine into a graphics controller converting it into a GPU was done to solve a geometry problem. Therefore, you will find a lot of discussion about geometry but no math in this book.

Transformation and lighting are critical components of computer graphics and the GPU. Transform means to convert the coordinates of the 3D model to the coordinates of the viewing or display device. The coordinates are described by the vertices of objects in a scene. Lighting refers to the simulation of light in a scene—on the objects in a scene—and the effect of light from one object to another and the scene. It is a complicated process.

As will be explained in later chapters, the GPU will find its way into general compute applications as a parallel processor, totally devoid of any graphics functions.

Graphics processing units were developed for three-dimensional (3D) computer graphics applications, first for CAD and then for games. The historical development of computer graphics gives an overview of the influences that heralded the development and invention of the GPU. The following is an overview of those developments, which will provide a foundation for appreciating the GPU’s development and evolution.

## Displays and Pixels

In computer graphics and digital imaging, a pixel is the smallest addressable element in a raster image, or the smallest addressable element in a digital display such as an LCD. A pixel is not addressable in geometry. Geometry is at the front end of a processing pipeline and the output is a raster scan, sized to a specific physical display, and measured in pixels.

A GPU’s primary function is to drive a display (although GPUs used for compute acceleration do not drive a display). GPUs and graphics controllers before them