

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 9225744

**The dynamics of the computer industry: Modeling the supply of
workstations and their components**

Touma, Walid Rachid, Ph.D.

The University of Texas at Austin, 1992

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

**THE DYNAMICS OF THE COMPUTER INDUSTRY:
MODELING THE SUPPLY OF WORKSTATIONS
AND THEIR COMPONENTS**

by

WALID RACHID TOUMA, B.S.E.E., M.S.E.E.

DISSERTATION

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

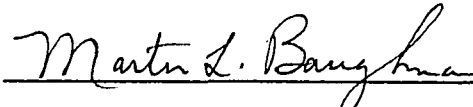
DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

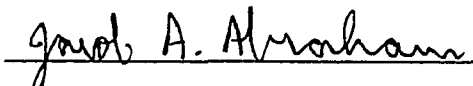
May, 1992

**THE DYNAMICS OF THE COMPUTER INDUSTRY:
MODELING THE SUPPLY OF WORKSTATIONS
AND THEIR COMPONENTS**

**APPROVED BY
DISSERTATION COMMITTEE:**




Martin L. Baughman, Supervisor



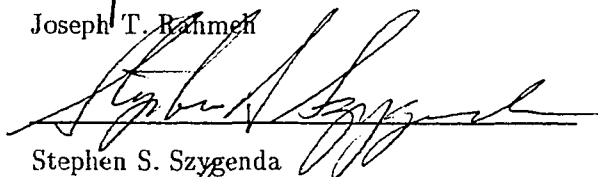
Jacob A. Abraham



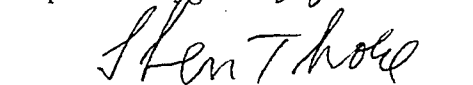
David A. Kendrick



Joseph T. Rahmel



Stephen S. Szygenda



Sten A. Thore

Copyright
by
Walid Rachid Tuma
1992

To the **WAR CHILD.**

Acknowledgments

Father Rachid, I love you and I respect you. You are my soul friend, you are my mentor in life and my inspiration, and I would not be who I am today without your emotional, mental, and moral support. Mother Laure, I love you and I kneel to the effort you have put into rearing me. I only wish I could have spent more time with you in my adult life so that I could have gotten to know you better. I guess the war has its prices.

Marty Baughman, I thank you and I sincerely acknowledge your mentorship and your partnership. We met in 1985 and, since then, we have had a genuine friendship, coupled with mutual respect and appreciation for each other's talents. Our friendship came to life as we teamed up on the "Workstation Project" and developed what has become, so far, a new research area that deals with understanding the dynamics of one of the most fascinating industries of our time: the computer industry. As our endeavors progressed, several new members joined the team, such as Jacob Abraham, David Kendrick, Joe Rahmeh, Steve Szygenda, and Sten Thore, my acknowledgements of whom will be presented later.

Katherine Gregory, my soul mate, my music, my metamorphosis, my woman: you have been the pillar of emotional support for me while I wrote this dissertation. Here's to you, Kathy G., to the wonderful, loving, and innocent child in you, to the mother in you, and to your soul.

Brothers Tanos and Jihad and your families, and Sister Nadine, my

best friends, my ears, my partners, and my support, I love you. Uncle Fadlo, my wise ear away from home, I love you, I sincerely appreciate the advice you have given me, and I thank you for assisting me in coming to the United States and in joining the University of Texas at Austin.

Jacob Abraham, I appreciate your support for the project since its inception. David Kendrick, I thank you for your wisdom, your honesty, and your patience. Joe Rahmeh, I thank you for all your support throughout my graduate career. Steve Szygenda, I thank you for your constructive criticism throughout the project. Sten Thore, at times, you have been like a father to me, and I am deeply grateful to your generosity and genuine soul.

Last but not least, I would like to thank everybody at the departments of Electrical and Computer Engineering, Economics, Mathematics, and Business, at the Computer Engineering Research Center, at the Center for Economic Research, and at the IC² Institute who assisted the team and me throughout the “Workstation Project” and while I wrote this dissertation.

Walid Rachid Touma

The University of Texas at Austin

May, 1992

**THE DYNAMICS OF THE COMPUTER INDUSTRY:
MODELING THE SUPPLY OF WORKSTATIONS
AND THEIR COMPONENTS**

Publication No. _____

Walid Rachid Touma, Ph.D.

The University of Texas at Austin, 1992

Supervisor: Martin L. Baughman

Discrete event simulation models are developed to capture the dynamics of supply of workstation hardware and software components. The models are tuned so that the results match actual trends of behavior for certain key attributes between 1980 and 1991. The tuned supply models are then used to project the costs of the workstation components from 1992 through 1996. The projected results are also used as input parameters to a linear workstation assembly model. The results of the workstation assembly model show that a top-of-the-line workstation, configured with a 200 megaHertz CPU, a 64 megabyte main memory, a 1 gigabyte magnetic hard disk, a 19-inch color CRT display with a 2.6 megapixel resolution, and a UNIX operating system, will cost around \$10,000 by 1994. These same capabilities, if configured into a workstation in 1991, would cost approximately \$20,000.

The models are used to perform sensitivity analyses with respect to the ICs' feature size and the number of silicon wafer defects per unit area. Three feature size cases are considered. In the Base Case, the feature size decreases at an exponential rate of 5.5% per year. In Case 2, the feature size stops decreasing after 1992, and the price of an assembled workstation jumps 20.5% from the projected Base Case price for 1996. In Case 3, the feature size decreases at an exponential rate of 10% per year—almost double the Base Case rate—and the price of an assembled workstation decreases 8.4% from the projected Base Case price for 1996. If extended to the year 2001, the 10% exponential rate of decrease of feature size yields a price per megabyte of a semiconductor DRAM that is less than the price per megabyte of a magnetic hard disk. This suggests that, if nonvolatile semiconductor DRAMs are developed by 2001, DRAMs could change the overall configuration of a computer by replacing the magnetic hard disk as the permanent storage component and pave the way to real-time computing.

Correspondingly, three values of the number of silicon wafer defects per unit area are considered. In the Base Case, the number of silicon wafer defects per unit area is 2.5 defects per cm^2 . In Case 2, the number of silicon wafer defects per unit area doubles after 1992 to 5 defects per cm^2 , the IC die yields decrease by as much as 88.9%, and the CPU price per megaHertz and the DRAM price per megabyte jump 277% and 69.1% respectively from their projected Base Case values for 1996. In Case 3, the number of silicon wafer defects per unit area is halved to 1.25 defects per cm^2 after 1992, the IC die yields increase by as much as 355.6%, and the CPU price per megaHertz and the DRAM price per megabyte decrease 51.8% and 47.7% respectively from their projected Base Case values for 1996.

Table of Contents

Acknowledgments	v
Abstract	vii
Table of Contents	ix
List of Figures	xvi
List of Tables	xix
1. Introduction	1
1.1 Problem Statement	1
1.2 Computer Technologies	1
1.3 Computer Systems and Workstations	2
1.4 Contributions	4
1.5 Key Findings	5
1.6 Outline and Summary	7
2. Definitions, Terminology, and Concepts	9
2.1 The Workstation	10
2.1.1 The Birth of the Workstation	10
2.1.2 The Workstation and its Components	11
2.1.3 The Workstation and its Attributes	17
2.1.4 Workstation Performance	17

2.2	Semiconductor ICs	24
2.2.1	Semiconductor Physical Characteristics	25
2.2.2	<i>pn</i> -Junction	26
2.2.3	Transistor Fabrication Technologies	27
2.2.4	High Speed IC Technologies	31
2.2.5	ICs: Speed versus Die Area	34
2.2.6	Reliability of Computer Systems	36
2.3	Magnetic Hard Disks	36
2.3.1	Computer Storage System	37
2.3.2	Magnetic Storage Technology	38
2.3.3	Magnetic Hard Disk Components	39
2.3.4	Magnetic Hard Disk Performance	47
2.3.5	Concluding Remarks	50
2.4	Color CRT Displays	51
2.4.1	CRT History	52
2.4.2	Competing Display Technologies	53
2.4.3	Color CRT Components	53
2.4.4	Screen Image	60
2.4.5	CRT Bandwidth	62
2.4.6	CRT Hardware Drivers	62
2.4.7	CRT Resolution	64
2.4.8	Concluding Remarks	66
2.5	UNIX Operating System	67
2.5.1	UNIX History	68
2.5.2	UNIX Design Structure	69

2.5.3	Stepping through a UNIX Command	71
2.5.4	Main UNIX Managed Functions	72
2.5.5	Concluding Remarks	74
3.	Workstation Supply Models	75
3.1	Simulation Overview	75
3.1.1	The Simulation Modeling Approach	75
3.1.2	Relational Diagrams and Attributes	78
3.2	ICs Supply Model: Microprocessors and DRAMs	83
3.2.1	Historical Data on the Physical Characteristics of ICs . .	84
3.2.2	Historical Data on ICs Capabilities and Price Trends . .	94
3.2.3	Model Assumptions and Terminology	104
3.2.4	CPU Speed and MIPS Models	107
3.2.5	DRAM Capacity Model	111
3.2.6	Die Areas for Fixed Capability ICs	113
3.2.7	IC Cost Model	114
3.3	Magnetic Hard Disk Supply Model	123
3.3.1	Historical Data on the Physical Characteristics of Mag- netic Hard Disks	124
3.3.2	Historical Data on Magnetic Hard Disk Capabilities and Price Trends	130
3.3.3	Model Assumptions and Terminology	134
3.3.4	Magnetic Storage Radius	137
3.3.5	Number of Disk Recording Tracks	137
3.3.6	Track Capacity and Density	139

3.3.7	Hard Disk Data Rate	140
3.3.8	Areal Capacity and Density	142
3.3.9	Volumetric Capacity and Density	143
3.3.10	Magnetic Hard Disk Cost/MB	143
3.3.11	Magnetic Hard Disk Total Cost	146
3.4	Color CRT Display Supply Model	147
3.4.1	Historical Data on the Physical Characteristics of Color CRT Displays	148
3.4.2	Historical Data on Color CRT Display Capabilities and Price Trends	150
3.4.3	Model Assumptions and Terminology	155
3.4.4	Bandwidth	157
3.4.5	Resolution	157
3.4.6	Metal Shadow Mask Manufacturing Yield	161
3.4.7	Color CRT Display Cost	162
3.5	UNIX Operating System Supply Model	166
3.5.1	UNIX Development-from-Scratch and Porting Trends . .	167
3.5.2	Model Assumptions and Terminology	169
3.5.3	UNIX Development-from-Scratch Time Period	171
3.5.4	UNIX Porting Time Period	174
3.5.5	Software Attributes Index	174
3.5.6	Workstation Hardware Attributes Index	176
3.5.7	UNIX Development-from-Scratch and Porting Costs . . .	177
3.6	Workstation Assembly Model	180
3.6.1	Assembly Steps	183

3.6.2	Model Assumptions and Terminology	183
3.6.3	Model Formulation	185
4.	Model Behavior and Sensitivity Results	188
4.1	Component Cost, Single Unit Price, and Bulk Price	190
4.2	Component Supply Model Inputs	191
4.2.1	Model Input Parameters	192
4.2.2	Components' Physical Characteristics Trends	193
4.2.3	Components' Capabilities and Price Trends	194
4.3	ICs: Model Results and Actual Market Data	194
4.3.1	ICs Die Yields	195
4.3.2	ICs: CISC CPUs Model Results and Actual Market Data	195
4.3.3	ICs: RISC CPUs Model Results and Actual Market Data	199
4.3.4	ICs: DRAMs Model Results and Actual Market Data . .	205
4.4	Magnetic Storage: Model Results and Actual Market Data . . .	207
4.4.1	Magnetic Hard Disk Price/MB	210
4.4.2	Magnetic Hard Disk Areal Density	210
4.4.3	Notes on Volumetric Density and Data Rate	213
4.5	Color CRT Display: Model Results and Actual Market Data . .	213
4.5.1	Color CRT Price/Megapixel	214
4.5.2	Color CRT Number of Pixels/Inch	214
4.6	UNIX: Model Results and Actual Market Data	217
4.6.1	UNIX Porting Times and Costs	217
4.6.2	UNIX Development-from-Scratch Time and Cost	219
4.7	Workstation Assembly Model: Inputs and Projected Results . .	219

4.7.1	Projected Results	223
4.8	Sensitivity Analyses	225
4.8.1	Sensitivity to the Feature Size	227
4.8.2	Sensitivity to the Number of Silicon Wafer Defects per Unit Area	241
5.	Conclusions and Suggestions for Future Research	249
5.1	Suggestions for Future Research	251
5.2	Final Comments	254
A.	Samples of the Supply Models Outputs	255
A.1	CPU Supply Model Output	256
A.1.1	CISC CPU MIPS, Speed and Cost Versus Die Size: 1980 - 1990	256
A.1.2	CPU MIPS, Speed and Cost Versus Year: 0.0625 cm ² - 4.1209 cm ²	262
A.1.3	Fixed Speed CPU Cost and Die Area: 1985 - 1990	267
A.2	DRAM Supply Model Output	269
A.2.1	DRAM Capacity and Cost Versus Die Size: 1980 - 1990	269
A.2.2	DRAM Capacity and Cost Versus Year: 0.0625 cm ² - 4.1209 cm ²	275
A.2.3	Fixed Capacity DRAM Cost and Die Area: 1985 -1990	280
A.3	Magnetic Hard Disk Supply Model Output	282
A.3.1	Magnetic Hard Disk Cost/MB	283
A.3.2	Magnetic Hard Disk Areal Density	284
A.3.3	Magnetic Hard Disk Parameters for Different Diameter and Height Specifications	284

A.4	Color CRT Display Supply Model Output	289
A.4.1	Maximum Number of Holes in the Metal Shadow Mask: 1985-1990	289
A.4.2	16-inch Color CRT Parameters and Cost: 1985-1990 . . .	290
A.4.3	19-inch Color CRT Parameters and Cost: 1985-1990 . . .	291
A.4.4	20-inch Color CRT Parameters and Cost: 1985-1990 . . .	292
A.4.5	25-inch Color CRT Parameters and Cost: 1985-1990 . . .	293
A.5	UNIX Supply Model Output	294
A.5.1	UNIX Porting Time Period and Cost: 1980-1990	294
A.5.2	UNIX Development-from-Scratch Time Period and Cost: 1980-1990	295

BIBLIOGRAPHY **296**

Vita

List of Figures

1.1	Computer classes versus user environment. Source: [4].	3
2.1	General computer system configuration.	12
2.2	<i>npn</i> bipolar transistor configuration.	28
2.3	Enhancement type NMOS transistor configuration.	30
2.4	CMOS inverter configuration.	32
2.5	Rigid disk file components. Source: [57].	40
2.6	Magnetic disk surface layout.	42
2.7	Inductive head structure.	44
2.8	CRT display components. Source: [90].	54
2.9	Inner screen phosphors layout structures. Source: [90].	59
2.10	UNIX layer structure.	70
3.1	Illustration of a relational diagram.	79
3.2	Die Sizes of CISC CPUs and DRAMs versus time. Source: [48].	86
3.3	Feature size versus time. Source: [92].	90
3.4	Silicon wafer diameter versus time. Source: [92].	93
3.5	Actual data on the number of instructions per cycle for Intel and Motorola CISC CPUs. Sources: Intel data [41, 42, 48], Motorola data [61, 71, 84].	100

3.6	Actual data on the number of instructions per cycle for HP-PA and Sun SPARC RISC CPUs. Sources: HP data [7, 22, 37, 51, 53, 89, 103], Sun data [84, 86].	101
3.7	Relational diagram of the operational speed of CPUs.	109
3.8	Relational diagram of the DRAM capacity.	112
3.9	Relational diagram of the average IC die testing time.	120
3.10	DASD recording system scaling. Source: [3].	126
3.11	Actual price/megabyte of magnetic hard disk storage versus time. Source: [88].	131
3.12	Areal densities of magnetic storage devices in bits/mm ² versus time. Source: [57].	133
3.13	Illustration of the dimensional parameters of a magnetic hard disk.	138
3.14	Relational diagram of the cost/megabyte of a magnetic hard disk.	144
3.15	Relational diagram of the number of pixels/inch of a color CRT display.	159
3.16	Relational diagram of the cost/megapixel of a color CRT display.	164
3.17	Relational diagram of the UNIX development-from-scratch time period.	172
3.18	Relational diagram of the UNIX porting time period.	175
3.19	Relational diagram of the workstation hardware attributes. . . .	178
3.20	Illustration of a workstation assembly network.	181

4.1	Present value prices of workstations: Types 1, 2, and 3.	224
4.2	Feature Size: Sensitivity of the price/megaHertz of CPUs - Cases 1, 2, and 3.	231
4.3	Feature Size: Sensitivity of the price/megabyte of DRAMs - Cases 1, 2, and 3.	232
4.4	Feature Size: Sensitivity of the price/megabyte of magnetic hard disks - Cases 1, 2, and 3.	234
4.5	Feature Size: Sensitivity of the price of a 19-inch color CRT dis- play - Cases 1, 2, and 3.	236
4.6	Feature Size: Sensitivity of the present value price of a type 2 workstation - Cases 1, 2, and 3.	238
4.7	Feature Size: Sensitivity of the DRAM and the magnetic hard disk prices/megabyte - Cases 1 and 3.	240
4.8	DPUA: Sensitivity of the price/megaHertz of CPUs - Cases 1, 2, and 3.	246
4.9	DPUA: Sensitivity of the price/megabyte of DRAMs - Cases 1, 2, and 3.	247

List of Tables

3.1	Die areas and their actual die yields. Source: [33].	88
3.2	Actual market data of Intel and Motorola CISC CPUs. Sources: Intel data [41, 42, 48], Motorola data [61, 71, 84].	95
3.3	Actual prices and price/MIPS market data of Intel CISC CPUs. Sources: [41, 42].	97
3.4	Actual performance data of HP-PA and Sun SPARC RISC ma- chines. Sources: HP data [7, 22, 37, 51, 53, 89, 103], Sun data [84, 86].	99
3.5	Actual market data on DRAM die sizes, capacities, and densities. Source: [48].	103
3.6	Actual prices and price/megabyte market data of Motorola DRAMs. Source: [62].	105
3.7	Actual market data of 19- and 20-inch color CRT displays. Sources: [1, 2, 16, 28, 38, 59, 75, 81, 83, 93].	151
3.8	Actual price/megapixel and number of pixels/inch market data of 19- and 20-inch color CRT displays. Sources: [1, 2, 16, 28, 38, 59, 75, 81, 83, 93].	154
4.1	Model results and actual market data of die yields for certain die areas in 1989. Source: [33].	196

4.2	Model results and actual market data on MIPS ratings and prices of Intel CISC CPUs. Sources: [41, 42, 48].	197
4.3	Model results and actual market data on the price/MIPS of Intel CISC CPUs. Sources: [41, 42].	200
4.4	Model results and actual market data on the speed/cm ² rating of CISC CPUs. Sources: Intel data [41, 42, 48], Motorola data [61, 71, 84].	201
4.5	Model results and actual market data on MIPS ratings and prices of HP RISC CPUs. Sources: [7, 22, 37, 51, 53, 89, 103].	203
4.6	Model results and actual market data on the speed/cm ² rating of RISC CPUs. Sources: HP data [7, 22, 37, 51, 53, 89, 103], Sun data [84, 86].	204
4.7	Model results and actual Motorola market data on DRAM capacities and prices. Sources: [48, 62].	206
4.8	Model results and actual Motorola market data on the price/megabyte of DRAMs. Sources: [48, 62].	208
4.9	Model results and actual Motorola market data on the number of DRAM megabytes/cm ² . Sources: [48, 62].	209
4.10	Model results and actual market data on the price/megabyte of magnetic hard disks. Source: [88].	211
4.11	Model results and actual market data on the areal density of magnetic hard disks. Source: [57].	212

4.12 Model results and actual market data on the price/megapixel of a 19-inch color CRT display. Sources: [1, 2, 16, 28, 38, 59, 75, 81, 83, 93].	215
4.13 Model results and actual market data on the number of pixels/inch of a color CRT. Sources: [1, 2, 16, 28, 38, 59, 75, 81, 83, 93].	216
4.14 Model results and actual market data on the UNIX porting time periods and costs. Source: [72].	218
4.15 Model results and actual market data on the UNIX development-from-scratch time periods and costs. Source: [72].	220
4.16 Projected prices of the workstation assembly components and raw materials. Sources: [33, 86].	222
4.17 DPUA: Sensitivity of the die yields for certain die areas - Cases 1, 2, and 3.	244

Chapter 1

Introduction

1.1 Problem Statement

The digital information age is here. Computers across the globe communicate via satellite or fiber optic links, wide area networks share resources thousands of miles away, and a child at home has access, at the press of a button, to a world of knowledge. Several technologies have made possible this computer era, driven it, and affected its dynamics over time. The problem to be addressed in this dissertation is the formulation of a model that interrelates the factors that drive the supply of these technologies over time to the attributes of the computers that are manufactured from them.

1.2 Computer Technologies

A computer system is a grouping of resources, hardware and software, accessed by application programs which conform to the computer's programming language. The hardware resources are the seven component assemblies of the computer system: the processing board, the memory board, the data storage system, the display monitor, the printer, the mouse, and the keyboard. Each hardware resource is in itself a collection of several components, the functions of which will be elaborated upon later. The software resources consist of two major components: the computer system's manager, referred to as the operating

system, and the applications programs.

The attributes of the above technologies, at any point in time, depend on a number of factors, dictated largely by the tradeoffs between product life, production yields, learning curves, the pace of technical change, competition, and the state of technology. These factors are interrelated here in deterministic simulation models. The behavior of each resource's supply is dynamically represented and then integrated with dynamic models of other resources to provide insight into the dynamics of the assembled computer products. The component supply models, to be presented in a later chapter, include:

- Integrated circuits (ICs) supply model: microprocessors and dynamic random access memories (DRAMs).
- A magnetic hard disk storage supply model.
- A color cathode-ray tube (CRT) display supply model.
- A UNIX operating system supply model.

1.3 Computer Systems and Workstations

Every computer system, once configured, has its own application specific attributes and its own market niche. Figure 1.1 illustrates this concept by relating the computer user environment on the horizontal axis to the computer use style on the vertical axis, and ranks the available computer systems with respect to those two criteria. For instance, the computer workstation is represented in this figure as a shared computer in a professional environment;

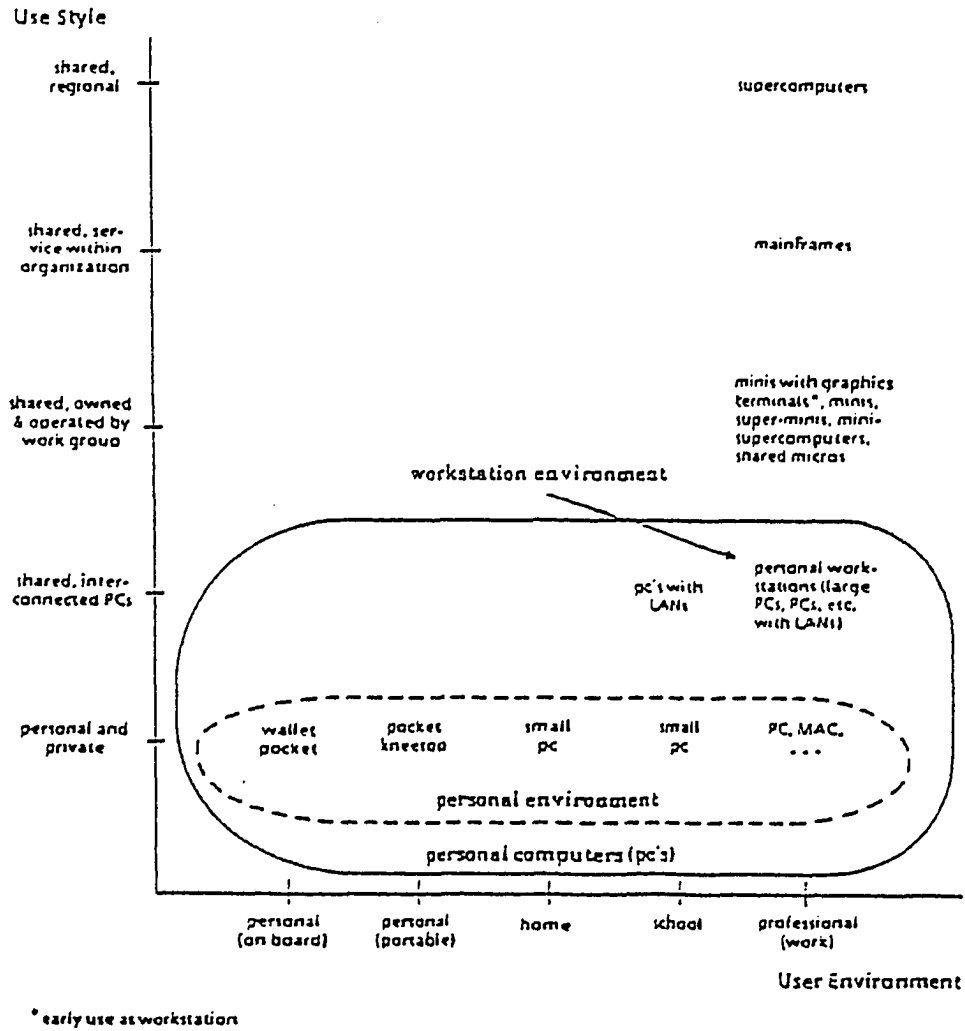


Figure 1.1: Computer classes versus user environment. Source: [4].

alternately, the personal computer (PC) is represented as a personal and private computer in a professional environment.

In the remainder of this dissertation, the computer workstation is focused upon because it combines the latest developments in processing power¹, DRAM capacities and densities, data storage system capacities and data rates, display resolutions, and computer management and computer applications software. Workstations, when interconnected to one another, form a single, shared (work and files) but distributed computing environment [4]. This concept of network computing and communication has catapulted the sales growth of the workstation higher than any technology in the computer industry, and sales are projected to grow at a compound rate of 30% per year over the next five years [85]. Due also to its high performance to price ratios [58], the computer workstation has emerged as the driving force in computing, a role formerly filled by mainframes and supercomputers.

1.4 Contributions

During the past four decades, the computer industry has emerged as one of the most—if not the most—dynamic industries since the industrial revolution. Never in history has there been a product with so short a lifetime, or period to obsolescence, compared to the time and effort that go into its research, development, and demonstration. One of the main objectives of this dissertation is to shed some light on the technology-driving trends of the computer industry

¹In this context, power is equivalent to speed and functionality.

and their effects on the dynamics of the industry as a whole, and in particular, on the supply of assembled workstations and their components.

The contributions of this dissertation to the state of knowledge are:

- Detailed analyses and documentation of the historical trends in the supply of hardware and software components used in a computer workstation.
- Deterministic discrete event simulation models that capture the past dynamics of various attributes of workstation components.
- A workstation assembly model that brings together all the supply models of the workstation components to provide insight into possible future behavior of the supply of fixed capabilities workstations.
- A tool to:
 - project the trends in the supply behavior of the industry for alternative scenarios of market or technological change.
 - perform sensitivity analyses with respect to certain technology barriers and their effect on the decision-making strategies of the companies supplying the technologies.

1.5 Key Findings

The key findings of this dissertation are:

- A \$10,000 price ceiling has become a standard for top-of-the-line workstations in today's distributed computing environment. By 1994, the model results from this dissertation show that a top-of-the-line workstation

will have the following hardware capabilities: a 200 megaHertz CPU, 64 megabytes of DRAM, a 1 gigabyte of magnetic storage, and a 19-inch color CRT display with a 2.6 megapixel resolution. These same capabilities, if configured into a workstation in 1991, would cost approximately \$20,000. This is a rate of decrease in price of over 20% per year for a fixed capabilities workstation.

- The feature size of integrated circuits is one of the most critical and influential technology-driving trends in the computer industry. The price results of all the workstation component supply models are sensitive to changes in the rate of decrease of the feature size over time. The effects of these changes are reflected in the overall price of an assembled workstation. In the Base Case, the feature size decreases at an exponential rate of 5.5% per year. In Case 2, the feature size stops decreasing after 1992, and the price of an assembled workstation jumps 20.5% from the projected Base Case price for 1996. In Case 3, the feature size decreases at an exponential rate of 10% per year—almost double the Base Case rate—and the price of an assembled workstation decreases 8.4% from the projected Base Case price for 1996.
- One of the possible consequences of accelerating the rate of decrease of the feature size is an accelerated decrease in the price per megabyte of semiconductor DRAMs. When the rate of decrease of the feature size doubles after 1992, the model results show that the price per megabyte of a semiconductor DRAM becomes cheaper than the price per megabyte of a magnetic hard disk by the year 2001, assuming also a continued fall

in magnetic disk drive prices. If nonvolatile semiconductor DRAMs are developed by 2001, DRAMs could change the overall configuration of a computer by replacing the magnetic hard disk as the permanent storage component and pave the way to real-time computing.

- A clean and precise ICs manufacturing environment and, consequently, a smaller number of silicon wafer defects per unit area greatly influences the reliability of the chips, their yields, and, ultimately, their costs. In the Base Case, the number of silicon wafer defects per unit area is 2.5 defects per cm^2 . In Case 2, the number of silicon wafer defects per unit area doubles after 1992 to 5 defects per cm^2 , the IC die yields from the model decrease by as much as 88.9%, and the CPU price per megaHertz and the DRAM price per megabyte jump 277% and 69.1% respectively from their projected Base Case values for 1996. In Case 3, the number of silicon wafer defects per unit area is halved to 1.25 defects per cm^2 after 1992, the IC die yields increase by as much as 355.6%, and the CPU price per megaHertz and the DRAM price per megabyte decrease 51.8% and 47.7% respectively from their projected Base Case values for 1996.

1.6 Outline and Summary

This dissertation is organized as follows:

- Chapter 2 presents some of the main definitions, terminology, and concepts related to computer workstation technologies.

- Chapter 3 presents discrete event simulation supply models of microprocessors, DRAMs, magnetic hard disks, color CRT displays, and UNIX operating systems, and a linear workstation assembly process model.
- Chapter 4 presents simulation results from the models. It compares the results of the component supply models with historical trends, presents the results of the workstation assembly model for three different workstation configurations, and analyzes the sensitivity of the component supply and workstation assembly models to variations in the projected rate of decrease of the IC feature size and in the number of silicon wafer defects per unit area.
- Chapter 5 presents the conclusions and suggests future research directions.
- Appendix A lists, in a series of tables, sample results of the CPU, DRAM, magnetic hard disk, color CRT display, and UNIX operating system supply models.

Chapter 2

Definitions, Terminology, and Concepts

This dissertation relies upon concepts relating to both computer industry technologies and mathematical modeling. These concepts are merged to create models that capture the dynamics of the supply of workstations and their components. This chapter provides some of the essential definitions, terminology, and concepts related to computer technologies. The mathematical modeling concepts will be discussed in detail in a later chapter.

- Section 2.1 presents an overview of the workstation and its attributes.
- Section 2.2 describes the manufacturing processes and attributes of semiconductor ICs.
- Section 2.3 describes the assembly and attributes of magnetic hard disks.
- Section 2.4 describes the assembly and attributes of color CRT displays.
- Section 2.5 presents an overview of the design and functionality of the UNIX operating system.

Volumes could be written about each of these technologies and its attributes. Only the basics are presented here. References are provided for readers interested in more complete descriptions.

2.1 The Workstation

Since the workstation is the computer system under consideration here, the following subsection presents a brief history of the emergence of the workstation. In Subsection 2.1.2, a description of the workstation component assemblies is presented. In Subsection 2.1.3 a workstation attribute is defined, and in Subsection 2.1.4 the factors affecting the workstation's performance and how to measure them are reported.

2.1.1 The Birth of the Workstation

The computer industry can be compartmentalized into user applications and computer system attributes that match these applications. The system attributes depend on the capabilities and specifications of the system's hardware and software components. Such capabilities will differentiate the system from other computers in price and performance. From the point of view of the user familiar with personal computers (PCs), the workstation is an advanced PC; i.e., a workstation has more processing power, more main memory, more data storage space, a better user interface, and more sophisticated display capabilities than a PC. Workstations edged out other forms of computing when they were introduced in 1980 with their networking capabilities, which are handled by the operating system or the software manager of the machine [4, 29]. Chief among the technical developments that led to the birth of the workstation were:

- The development of an operating system that handled distributed and multitasking computing environments. The Digital Equipment Corporation (DEC) and Xerox were the leading developers of such software [4].

- The commercialization of the local area network (LAN) software and the Ethernet hardware in 1981.
- Breakthroughs in the semiconductor industry, especially the availability of cheaper and denser memory chips and a more powerful array of microprocessors and microcontrollers.
- The increase in the disk¹ storage density and its data-access rates.
- Availability of displays with over 1 million pixels, driven by sophisticated hardware that enabled higher refresh rates and better resolutions.

2.1.2 The Workstation and its Components

All workstations have a common set of component technologies. Illustrated in Figure 2.1 are the three major hardware component assemblies and the software component assembly. The hardware component assemblies include the processing board, the memory board, and the external input/output (I/O) interfaces (keyboard, mouse, data storage, display, and printer). The software component assembly includes the system resources manager—operating system—and the computer/user window-interface manager². Each hardware component assembly is shown in a bold-typed rectangle and connected to another hardware assembly/ies by a bold-typed line. The software component assembly is shown

¹A disk, in this context, is a data storage system that uses either magnetic or optical technologies.

²The system resources manager and the computer/user window-interface manager can reside permanently in a storage device (magnetic or optical hard disk), or temporarily in main memory while the computer system is powered on.

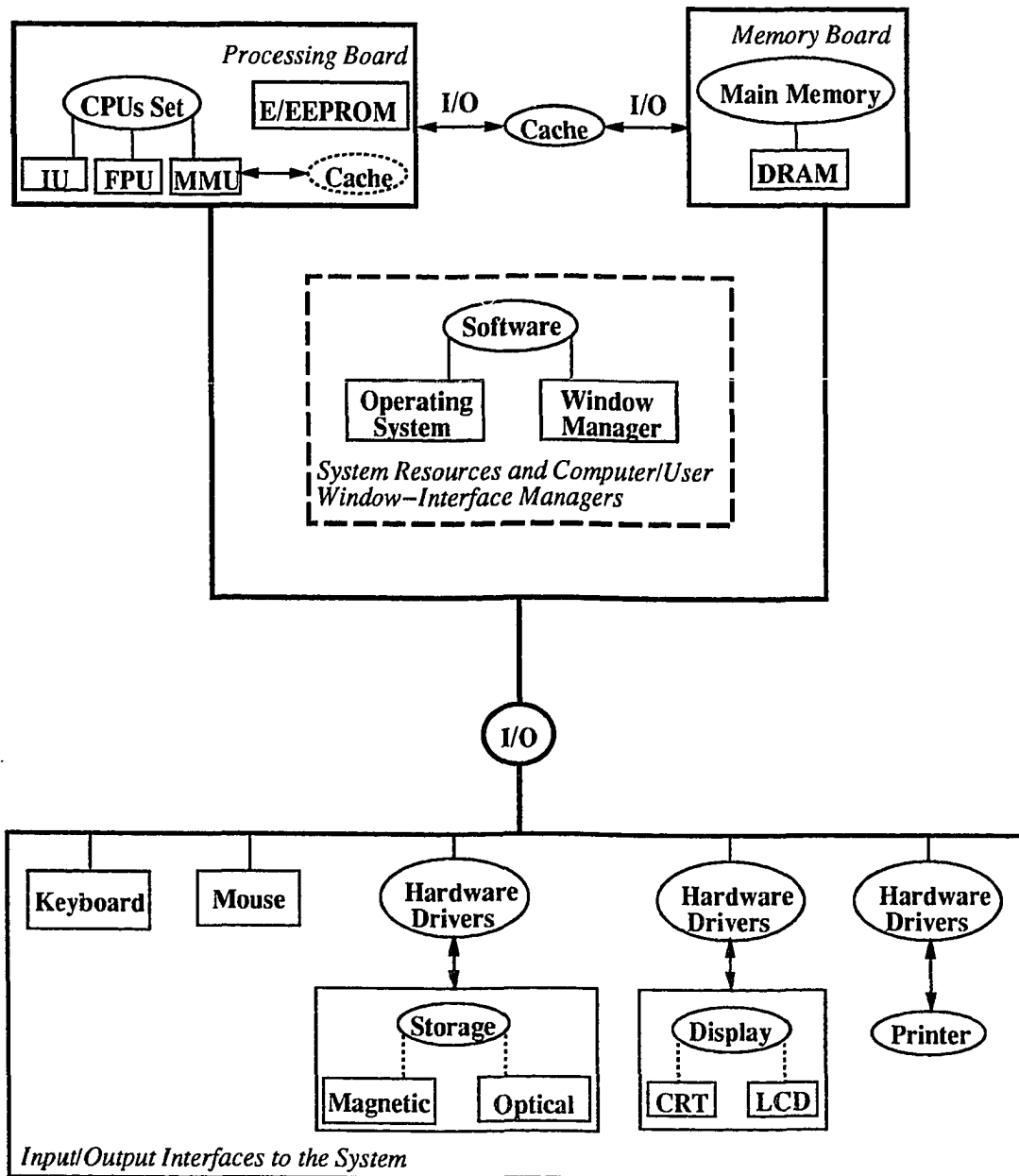


Figure 2.1: General computer system configuration.

in a broken rectangle to indicate that it is not a hardware component, but acts as a hardware component assemblies manager. Some components in the hardware assemblies are shown in broken circles to illustrate that their presence is optional, like the cache³ in the processing board assembly. In the case of the storage technologies, the magnetic and the optical boxes are connected to the storage element of Figure 2.1 with broken lines to indicate that either could be used as the storage technology in the workstation. Similarly, the CRT and the LCD⁴ boxes are connected to the display element of Figure 2.1 to indicate that either could be used as the display technology in the workstation.

The hardware component assemblies are connected in an hierarchical order: first, the processing and the memory boards where the task execution is performed; and second, the I/O interfaces where the task can be keyed in, stored, displayed, or printed. The software component is shown in Figure 2.1 between the two hardware layers of the computer system to indicate that it manages the execution of the task and its interaction with the I/O interfaces. The two hardware layers are connected by communication lines, the collection of which is called the bus. A bus can be 8 bits⁵ to 128 bits wide, depending on the architectural configuration of the microprocessor, the main memory⁶, and the I/O interfaces.

³If the machine price is not an issue, optional cache memories can be installed on the main memory board and on hardware driver boards of the I/O interfaces.

⁴LCD is equivalent to liquid crystal display.

⁵The bit is the digital representation of information. Its values are binary (0 or 1), and a series of eight bits represents a byte.

⁶Main memory can be considered as an I/O interface to the microprocessor, but in the hierarchy adopted, it is a part of the processing team.

What follows is a brief description of the component assemblies in Figure 2.1 and their corresponding subcomponents:

- The processing board includes an integer unit, a floating point unit, a memory management unit, and, in some cases, the cache memory and/or read-only-memory (ROM) chips. The integer and floating point units handle several types of instructions such as arithmetic, logical, control, floating point, or data transfer operations. The memory management unit handles the interfaces between the processing units and the cache memory, between the cache memory and the main memory, and between the main memory and the disk memory. The cache is a static random access memory (SRAM) sandwiched between the processing units and the main memory to reduce the wait-on-data cycles of the processors. Usually, the cache is the fastest type of memory used in the workstation and has the highest cost per byte of memory used. The ROMs⁷ contain subroutines that perform specific tasks such as computer startup and initialization procedures.
- The main memory board consists of DRAMs. The application program and the data it manipulates are located in DRAMs during the program's execution. An application program is a sequence of instructions that perform a task. An instruction specifies the arithmetic and logic operations to be executed and governs the transfer of information within the computer's resources. Data in computers is any digital information representing numbers and encoded characters to be used during the program execution;

⁷ROMs are usually configured as electrically erasable and programmable read-only-memories (EEPROMs).

data can be entered in the computer system via the keyboard or loaded into main memory via a backup storage device, such as a magnetic hard disk, tape, or diskette. Once the program is in main memory and ready to be executed, the processor fetches the instructions and performs the desired operations. When the program finishes execution, the results are either displayed on the monitor or stored onto disk for later reference. For a more elaborate description of the process of executing a program, see Hamacher/Vranesic/Zaky [31] and Hennessy/Patterson [33] .

- The data storage system consists of magnetic or optical disks. These disks retain the digitized information indefinitely, unlike the DRAMs where the data disappears as soon as the machine's power is turned off. Disk data rates (megabytes/second) and areal densities (bits/inch²) affect the efficiency of the computer system and the types of applications it can perform. The communication between the hard disk and the other computer resources is the most time consuming operation and, to reduce the resulting communication delays, several cache systems have been implemented between the disk and the computer's main memory. If the data requested by a program in execution is not in main memory, the processor has to stall⁸, and the execution is stopped until the disk comes back with the data. Hence, a trend in computer systems design will provide enough main memory⁹ in the machine so that the program and most of its data are

⁸In multitasking machines, the processor could perform another task while waiting for the data from the disk.

⁹assuming dense and cheap DRAMs are available

loaded in main memory before the execution starts, minimizing—if not eliminating—the references to the disk during execution.

- The high definition display consists of a color or black and white (B&W) cathode-ray tube (CRT) or liquid crystal display (LCD), and the necessary image-driving hardware. The image drivers are fast input/output (I/O) chips and play an important role in determining the display's resolution, refresh rates, and color templates. To the user, the display in a computer system is the main tool for displaying the output of an application program. Printers are also output devices, but they are too slow, expensive, and impractical for the highly interactive and animated software available on the market today; their best use is in obtaining a hard copy of finished work.
- For the past decade, UNIX has been the most effective operating system for workstations. UNIX manages the workstation's resources and coordinates the interface between the user and the machine and between the user and other network workstations. Of utmost importance to the workstation environment is its ability to network with other machines. This capability is supported by the UNIX distributed environment which provides the platform and the communication protocols for building large and wide area networks (LANs and WANs) of workstations. Several "window"-based software applications have been developed in the 1980s to manage and facilitate the computer/user interface. One of the most popular "window"-based systems running on top of UNIX in computer workstations is the X

windows manager¹⁰.

2.1.3 The Workstation and its Attributes

A workstation attribute is a measurable characteristic that enhances the workstation's value and reflects the sophistication of its technology. Each of the workstation components presented in Subsection 2.1.2 has its own specifications and adds enhancements to the operation of the workstation as a whole. Of the many workstation attributes, the main ones are the processing board's speed, the amount of DRAM, the storage system's capacity and response time, the display's resolution and color, the operating system's user friendliness, networking, and multitasking capabilities, and, most importantly, the workstation's total cost. All of the previously mentioned attributes¹¹ come together to influence the workstation's overall performance attribute, presented in detail in the next subsection.

2.1.4 Workstation Performance

The performance of a workstation is measured by the time a program takes to finish its execution. The execution time is measured by the following equation:

$$Time_{exec} = IPP * CPI_{ave} * CT \text{ (sec)} \quad (2.1)$$

¹⁰The X windows manager was developed at the Massachusetts Institute of Technology (MIT) with the support of the Digital Equipment Corporation (DEC); currently, an X consortium is handling the system's updates and enhancements.

¹¹Attributes such as workstation size and weight are not discussed here because they require the study of workstation demand attributes and this dissertation is concerned with supply attributes only.

where IPP is the number of instructions per program, CPI_{ave} is the average number of machine clock¹² cycles per instruction, and CT is the clock cycle time in seconds. The designer of a high performance machine reduces the average execution time of a program by effectively reducing the factors contained in the right-hand side of Equation 2.1. The elements affecting the right-hand side factors of Equation 2.1 and a description of the workstation performance measurement techniques are presented below.

IPP: The Number of Instructions per Program. The number of instructions per program depends on the compiler and the architecture of the machine. A compiler maps the source code written in a high-level language, such as C or Fortran, onto the machine's hardware language or assembly code. The most common machine architectures are CISC¹³ and RISC¹⁴.

The CISC instruction set is complex because most instructions require several minor operations during execution. The instructions have varying lengths in bits¹⁵, making the decoding¹⁶ circuitry and the memory access procedures more elaborate and complicated. A CISC instruction, on average, takes more than one cycle to terminate; nevertheless, the number of CISC instructions per compiled program is small compared to a RISC compilation.

¹²The machine clock is the computer cycle time or the inverse of the frequency at which the computer is running.

¹³CISC is an acronym for Complex Instruction Set Computing.

¹⁴RISC is an acronym for Reduced Instruction Set Computing.

¹⁵In a digital machine, data and instructions are represented by bits.

¹⁶Decoding is performed on a fetched instruction from main memory into the processor to determine the procedures the processor must follow to execute the fetched instruction.

The initial objective of **RISC** architecture was to have a set of simple one-cycle instructions [15]. Several scientists and institutions worked on the development of the RISC architecture, pioneered by IBM with the 801 minicomputer project [15], by David Patterson with the RISC¹⁷ project at the University of California at Berkeley, and by John Hennessy with the MIPS project at Stanford University. All RISC instructions have the same size—32 bits; most of them require one cycle to terminate, except floating point instructions which require more than one cycle. Complex floating point operations are usually software-implemented or rerouted to a dedicated floating point processor. Memory access instructions in RISC are very simple, composed of basic load and store commands. Such simplicity in the instructions set induces a simpler hardware design than CISC's and, consequently, faster machine operational speeds. Even though the RISC instruction set is simple, it takes several RISC instructions to perform comparably to one CISC instruction. There are tradeoffs, then, with both architectures: a CISC program has fewer instructions than a RISC, and a RISC instruction requires fewer computer cycles to terminate than does a CISC.

Once the instruction set is chosen, an implementation technique for reducing the number of instructions per program is to optimize the **compiler's** source code¹⁸ to object code¹⁹ mapping techniques [15, 34]. Compiler technologies flourished with the availability of cheap and dense DRAMs and the increase in the demand for RISC-based workstations. When main memory became less expensive, machines with greater amounts of DRAMs were affordable, and larger

¹⁷The first RISC microprocessor, RISC 1, was designed in 1982.

¹⁸Source code is equivalent to high level code like C.

¹⁹Object code is equivalent to machine language or assembly code.

programs with more instructions could reside in them. And, since a RISC compiled program has a great number of instructions, efficient mapping techniques became the focus of RISC compiler developers to compete with the speed of execution of a CISC. RISC compilers were designed to optimize the distribution of the code within the constraints of the computer's architecture, even if the architecture was not the most suitable for the application [82].

CPI_{ave}: The Average Number of Clock Cycles per Instruction. Several hardware organization and implementation factors influence a computer's average number of clock cycles per instruction. The most important ones are instruction pipelining and memory caching [34]. Instruction pipelining²⁰ is an architectural paradigm that improves the throughput of the machine without changing the basic cycle time by paralleling instruction executions in one processor. A pipeline can have several stages, where each stage gets dedicated to an instruction until it terminates. Since each instruction has several steps to be performed before termination, pipelining those steps can result in, on average, one instruction per computer clock cycle and, in some cases, two or more [71, 97], depending how many instructions are launched simultaneously per cycle and on the degree of sophistication of the hardware implementation.

The throughput can be increased by running in parallel more than one *functional unit*, with each unit performing a particular task²¹. In today's workstations, this form of parallelization is sometimes implemented through one

²⁰Instruction pipelining was introduced in mainframes in the early 1960s.

²¹The superscalar architecture of the IBM RISC System/6000 workstation implements this idea.

dedicated integer unit (IU), one floating point unit (FPU), and one memory management unit (MMU) per machine. These units may be integrated in one chip or separate, and most of their operations are hardware controlled. The form of parallelization which integrates more than one similar processing unit into one workstation²²—i.e., the machine is configured with several IUs, FPUs, and MMUs—has been developed [86], but cost reduction and software improvements are needed before such workstations gain any market share.

Memory caching is the most vital operation between the processor and the workstation's resources. Implemented inside or outside the processor, caches are the key tool in reducing the wait-on-data time (from main memory or disk) of a program during execution. Moreover, since the DRAM access speed increases at a much lower rate than do processors [34], more cache²³ memories are needed to fill that communication speed gap to accommodate the much faster operating rates of the processors.

CT: The Clock Cycle Time. The clock cycle time is the inverse of the hardware operating frequency (clock speed). Several workstations today operate at frequencies larger than 50 megaHertz (MHz), like the Hewlett-Packard (HP) 9000/730 (RISC) and the Intel i486 based machines (CISC) which operate at 66 megaHertz. It is safe to assume that speeds in the 200 to 250 megaHertz range will be attained before the end of the decade [64, 71, 97]. Nevertheless, several

²²A computer with more than one similar processing unit is called a parallel machine. Most parallel machines have either a SIMD (single instruction-multiple data) or a MIMD (multiple instructions-multiple data) configuration.

²³Since the costs of cache memories and their controllers are decreasing, caches have found their way into hard disk controllers and display and printer drivers.

physical barriers may become apparent when speeds reach the 0.5 to 1 gigaHertz (GHz) range, among them the wave reflection phenomenon. As the distance traveled by the electric signal decreases to accommodate higher frequencies, problems with wave reflections occur [82]. A wave reflection occurs when a signal generated by the line driver is received by a circuit whose impedance²⁴ is less than the line's impedance. To preserve Ohm's law²⁵, part of the signal travels back in the direction of the driver, and a wave is reflected; the noise generated by these reflections creates erroneous behavior in the whole machine. Other physical barriers are the fabrication of the ICs and their packaging. These will be described in later sections.

Workstation Performance Measurement. Measuring the performance of a workstation is an art [33, 43, 52, 98]. Using benchmark programs is the most common procedure to measure the performance of a workstation. SPEC²⁶ is the leading benchmark currently in use in the computer industry [98]. It is a collection of ten programs—four integer intensive and six floating point intensive programs. The performance of a machine is expressed as the geometric mean of the respective ratios of the execution time of the ten benchmark programs on the VAX 11/780 to their execution time on the benchmarked machine. Since most of the other popular benchmarks do not consider the system's configuration²⁷, load

²⁴The impedance is equal to the ratio of the phasor equivalent of a steady-state sine-wave voltage to the phasor equivalent of a steady-state sine-wave current, V/I .

²⁵ $Z_{ohms} = \text{impedance} = V/I$.

²⁶SPEC is an acronym for Systems Performance Evaluation Cooperative.

²⁷The configuration of the system is equivalent to the chosen processors and their speeds, the amount of DRAM, the capacity of the hard disk, and the type of operating system: each MIPS ought to be matched with 1 megabyte of DRAM and 1 megabit/second (Mb/s) throughput

distribution, or the size of the tasks run, the system's evaluator must incorporate their effects into the overall performance results before reporting them [43].

Metrics performance measures, like MIPS²⁸ or MFLOPS²⁹, have gained popularity since the emergence of RISC architectures, and they have been used to lure CISC users to RISC with the higher performance metrics obtained by RISC based workstations. Even though the MIPS and MFLOPS metrics do not report on the overall system throughput or response time which form the basis of the performance evaluations, the numbers are still interesting to computer engineers who want to capture a relative system performance and combine it with the system's speed to obtain the average number of instructions executed per machine clock cycle [43, 98]. MIPS can be expressed as:

$$MIPS = System\ SPEED_{MHz} * IPC_{ave} \quad (2.2)$$

$$= \frac{1}{CT * CPI_{ave}} \quad (2.3)$$

where $SPEED_{MHz}$ is the processor's speed in megaHertz and IPC_{ave} is the average number of instructions per clock cycle. Equation 2.2 shows that the MIPS metric is dependent on the machine's instruction set, making the comparison of MIPS ratings of machines with different instruction sets (CISC versus RISC) meaningless [33]. Furthermore, since the average number of instructions per cycle varies from one program to the other, the MIPS metric can vary on the same machine, making it a relative measure of performance and not an absolute.

of I/O [33].

²⁸MIPS is an acronym for Million Instructions Per Second.

²⁹MFLOPS is an acronym for Million Floating Point Operations Per Second.

2.2 Semiconductor ICs

Signal propagation is a major source of delay in a computer system. Smaller ICs and smaller boards result in smaller and faster machines, assuming the functionality of the ICs stays constant. Functionality reflects the degree of integration in a chip and the complexity of the tasks it can perform. For example, current microprocessors perform integer, floating point, and memory management operations, making their functionality far superior to microprocessors ten years ago in which integer operations were the norms of IC integration. Decreasing the size of an IC implies reducing the die size. A die, sometimes referred to as chip, is a small unpackaged functional element made by subdividing a wafer of semiconductor material³⁰. The wafer itself is laser-sliced out of a semiconductor ingot³¹, the manufacturing of which must be very uniform and clean to reduce the number of wafer defects and, consequently, reduce the costs of the dies.

This section discusses the technological and manufacturing trends for semiconductor ICs.

- Subsections 2.2.1 and 2.2.2 present the semiconductor physical characteristics and the *pn*-junction configuration, respectively.
- Subsection 2.2.3 describes the most utilized transistor fabrication technologies.

³⁰The definition of the word “die” was obtained from the IEEE Standard Dictionary of Electrical and Electronic Terms (IEEE-SDEET).

³¹A semiconductor ingot is cylinder shaped, with a diameter equal to the wafer diameter; the companies that manufacture the ingots must pass certain high quality tests before they can be considered as ingot suppliers [71].

- Subsection 2.2.4 presents three technologies for increasing the IC speeds of operation while keeping the same die area.
- Subsection 2.2.5 provides two ways of improving the IC speeds of operation by decreasing the die area.
- Subsection 2.2.6 briefly discusses the issues of computer system reliability and fault tolerance.

2.2.1 Semiconductor Physical Characteristics

Semiconductors are electronic materials with a range of resistivity between insulators and metals: at high temperatures, they behave as conductors, and at low temperatures, they behave as insulators. Silicon is the most widely used semiconductor; it is cheap and readily available in nature as sand. Other semiconductors like germanium (Ge) and gallium arsenide³² (GaAs) are used in the electronic industry, GaAs being an integral part of very fast switching circuits.

Silicon has a diamondlike crystal atomic structure [27]. It belongs to Group IV of the chemical periodic table; each atom has four valence electrons with other silicon atoms. The electron pairs of each crystal silicon atom are shared with other atoms with what is called a covalent bond. As the temperature of the silicon rises, electrons may be freed from the covalent bond and a hole created. A hole acts like a positive electron charge in a silicon crystal lattice where an electron is missing in one of the covalent bonds.

³²Gallium arsenide is not available in nature; it has to be formulated.

The process of hole and electron creation is referred to as ionization. An intrinsic semiconductor is a pure semiconductor, where the concentrations of holes and electrons are equal. An extrinsic semiconductor is a doped or impure semiconductor where holes or electrons are injected³³ in the silicon in order to create an imbalance in the charge concentrations. If the concentration of electrons is larger than the concentration of holes, the semiconductor is referred to as *n*-type or *n*-doped, and if the concentration of the holes is larger than the concentration of electrons, the semiconductor is referred to as *p*-type or *p*-doped.

2.2.2 *pn*-Junction

By joining a *p*-type and an *n*-type semiconductor or by diffusing *n*-type impurities into a *p*-type semiconductor or vice versa, a *pn*-junction is created. The *pn*-junction is the most important physical part and concept of the semiconductor electronic devices sector. It is at the core of the semiconductor transistor³⁴, the driving device of today's DRAMs, CPUs, and all the other application specific ICs (ASICs). For an historical background of the semiconductor industry, its economics, and the major players and their share of the world semiconductor markets, consult Yoffie [104].

³³Temperature increase can create a charge imbalance, but usually the semiconductor is operated at temperatures much lower than the ones that could create such an imbalance.

³⁴The first solid state transistor was developed at Bell Telephone Laboratories in 1947 by William B. Shockley and his team [104].

2.2.3 Transistor Fabrication Technologies

A transistor is an active semiconductor analog device with three or more terminals³⁵. Of the many transistor types utilized in ICs, the bipolar junction transistor (BJT) and the field-effect transistor (FET)—including CMOS³⁶, which is a FET-derived technology—are the most common. What follows is a description of the physical process of building these different transistor types.

Bipolar Junction Transistor. A bipolar junction transistor is formed by connecting two *pn*-junctions back to back. The configuration of a BJT can be either *pnp* or *npn* and, since the electrons have a higher mobility than the holes, the *npn* configured BJT is the most widely used in building ICs. Figure 2.2 presents the *npn* BJT configuration, where the *p* layer is referred to as the transistor's base and the outer and inner *n* layers are referred to as the transistor's collector and emitter, respectively. For more details on the BJT operation modes and circuit configurations, see Chapter 2 of Ghausi [27].

Field Effect Transistor. There are two basic types of FETs, junction and metal-oxide semiconductor FETs (JFETs and MOSFETs). The *pn*-junction and the electric field controlled current are the FETs' basic operational mechanisms. The MOSFET is the most widely used transistor in monolithic ICs³⁷ because

³⁵IEEE-SDEET

³⁶CMOS is equivalent to Complementary Metal-Oxide Semiconductor FET.

³⁷ICs are divided into two categories: monolithic ICs which are fabricated on a single semiconductor substrate and hybrid ICs where various components on separate chips are mounted on an insulating substrate and interconnected.

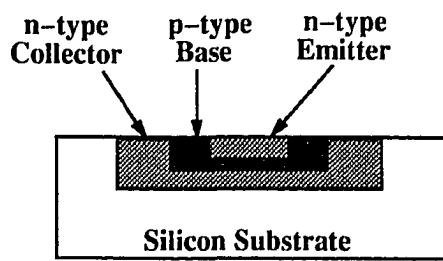


Figure 2.2: *npn* bipolar transistor configuration.

it occupies less space³⁸, has a high input impedance (i.e., draws less current and, eventually, consumes less power), has a low fabrication cost, and is less noisy than the BJT. There are two types of MOSFET, the enhancement and the depletion types, and each can have the main charge carriers as electrons or as holes. If the main carriers are holes, the MOSFET is called a *p*-channel MOSFET or PMOS, and if the main carriers are electrons, the MOSFET is called an *n*-channel MOSFET or NMOS. As mentioned earlier, electrons have a higher mobility than holes, so NMOS transistors are more frequently used as the building block of ICs.

Presented in Figure 2.3, an enhancement type NMOS transistor is built by diffusing two heavily doped *n*-type regions, referred to as source and drain, in a lightly doped *p*-type substrate. A silicon dioxide (SiO_2) layer covers the source and the drain, over which a metal plate, like aluminum or silicon, is deposited to form the FET's gate. To build a depletion type NMOS out of an enhancement type, a thin channel of lightly doped *n*-type material is diffused between the heavily doped *n*-type source and gate before covering them with the oxide layer. Of the two types, the enhancement is more widely used due to the simpler fabrication steps and the lower power consumption characteristics³⁹.

CMOS. Due to the low power consumption of enhancement MOSFETs and advancements in the IC manufacturing techniques, enhancement type PMOS

³⁸In general, MOS transistors occupy the least space in IC fabrication and, on average, the ratio of MOSFET space to BJT space is 0.2, or smaller by a factor of 5.

³⁹A depletion type MOSFET consumes more power than an enhancement type due to the presence of the thin channel between the source and the drain. The channel in a depletion type MOSFET gives rise to leakage currents under static or direct current (DC) voltage conditions.

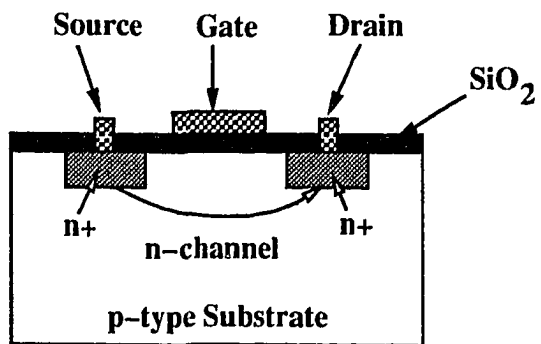


Figure 2.3: Enhancement type NMOS transistor configuration.

and NMOS transistors can be fabricated on the same IC to obtain the complementary symmetry MOS or CMOS⁴⁰ inverter (see Figure 2.4). For more details on the operation modes and the CMOS circuit configurations, see Chapter 3 of Ghausi [27]. Due to its temperature stability, lower power consumption than NMOS, low cost of production, and high packing density, CMOS has captured a high percentage of the very large scale integration (VLSI) sector of microelectronics [48, 50, 74, 100]. However, one major drawback of CMOS is its limited switching speeds. The maximum projected speed for CMOS based ICs is in the 70 to 100 MHz range [63, 71].

2.2.4 High Speed IC Technologies

For applications requiring higher circuit switching speeds, three alternatives exist: BiCMOS technology, GaAs technology, and optical technology.

BiCMOS. The BiCMOS fabrication technology is formed by combining the CMOS technology and the bipolar technology, known for its high switching speeds (100 - 200 MHz [63, 71]), high power consumption, and high fabrication costs. This alternative was driven by the bipolar technology manufacturers—who in the 1980s saw their market share slipping to CMOS—and by the proliferation of very fast cache memories (SRAMs) in microcomputers and workstations. New packaging techniques and materials and new cooling techniques are being developed to handle the high power consumption and the heat generated by BiCMOS ICs [48].

⁴⁰The CMOS configuration was invented by Frank Wanlass in 1963.

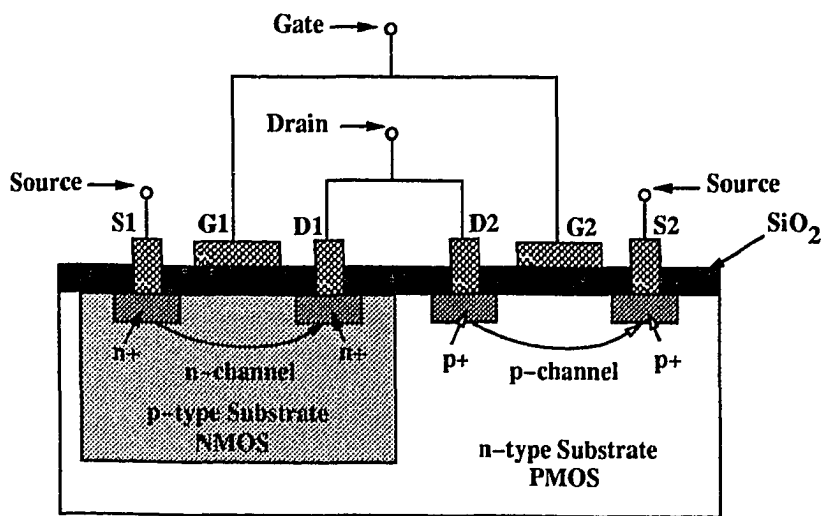


Figure 2.4: CMOS inverter configuration.

GaAs. This alternative uses gallium arsenide instead of silicon as the semiconductor. Gallium arsenide, formulated in the 1950s by Henry Welker of Siemens Laboratories, has an electron mobility of up to 6 times that of silicon with circuit switching speeds exceeding the 1 gigaHertz (GHz) range; it uses less power than silicon and can convert electronic signals to light [10, 12, 24, 27]. Furthermore, GaAs ICs costs, which were once considered so outrageous that only the DoD⁴¹ and supercomputer manufacturers could use them in their products, are decreasing significantly for three reasons [10, 12, 63]: 1) the GaAs manufacturers were able to map the silicon process technology to their GaAs ICs⁴² manufacturing techniques, 2) larger GaAs wafers are being produced, and 3) higher production yields are attained in the manufacturing laboratories. Another attractive feature of GaAs is that it can become superconductive if cooled to a -263 °C [24], which opens the door, previously barred by difficult manufacturing processes [36], to superconductivity and to superconductive ICs for integration in the supercomputers of the future⁴³.

Optical. The final alternative to increasing the ICs switching speeds is to use photonic, rather than electronic, transmission of signals. Photonic transmission involves optical interconnections and switches which can be obtained by laser diodes used as light transmitters and photo detectors [82]. Gallium arsenide, one of the prime materials in microwave circuits fabrication, can be used in

⁴¹DoD is an acronym for Department of Defense.

⁴²GaAs was mostly used in satellite microwave applications, converting and amplifying microwave signals in the gigaHertz range to electronic signals.

⁴³To keep supercomputers operating at normal temperatures, cooling processes are already being used to absorb the heat generated from their circuitry.

the manufacture of the laser diode. Although the optical technology switching speeds can reach the gigaHertz range, it is still a relatively new technology and prohibitively expensive for use inside a commercial computer. Optical computers may be available by the end of this decade or the beginning of the next, but the transition will undoubtedly not be easy [82].

2.2.5 ICs: Speed versus Die Area

After presenting an introduction to the semiconductor materials and technologies, it seems appropriate to discuss ways of increasing the speed of operation of a die by decreasing its area, while keeping the same functionality of the IC. There are basically two ways of accomplishing this, either by decreasing the feature size or by reconfiguring the IC layout.

Feature Size. As the feature size decreases, the ICs die areas decrease, making faster ICs switching speeds possible. The feature size is the minimum resolvable distance separating two etched lines in the semiconductor substrate. A line is etched in the semiconductor substrate by a process called lithography. Lithography is the process of transferring a circuit configuration on the semiconductor substrate. Of the many lithography methods, optical lithography is leading the electronics industry in the 1990s [21]. Optical lithography uses an ultra-violet beam to trace a pattern in photoresist. Photoresist is a substance that hardens when exposed to certain light frequencies, for example, ultra-violet. Developing the corresponding photoresist chemical structure for the particular wavelength of light has been a challenge to ICs manufacturers [21]. Nevertheless, when the wavelength of the light hitting the photoresist decreases, the feature size

decreases. Once the circuit pattern is transferred to the photoresist, the non-hardened part is scraped off the semiconductor and impurities are diffused in the silicon to create the doping characteristics described earlier. Details on the factors that influence the resolution in the photoresist and, consequently, the feature size of the etching process, can be found in [21] and Appendix B of Ghausi [27].

Optical lithography⁴⁴ can reach a minimum feature size of 0.2 micron-meter and, considering that a gigabit (Gb) DRAM⁴⁵ chip requires a feature size of 0.15 micron, optical lithography may very well lead the semiconductor industry to a 256 megabit, if not a 512 megabit DRAM by the end of the decade. The other lithography contenders, like X-Ray (with minimum feature size below 0.01 micron), electron beam, and ion beam (both with minimum feature sizes below 0.1 micron [21]), do not yet have the process maturity and high yields to be used in the production lines. Large companies like IBM and AT&T are pushing these technologies, hoping for a big payoff by the turn of the century [21].

Circuit Layout. Another way of decreasing the die area is by reconfiguring the layout of the IC and optimizing the number of connections among the different subcircuits embedded in the chip. But the layout problem and the min-cut

⁴⁴Lithography or, indirectly, feature size, accounts for two thirds of the increase in DRAM densities [21].

⁴⁵DRAMs are the test bed of any increase in feature size because they are the easiest chips to manufacture.

problem are NP-complete⁴⁶ [25, 47], making this option even more difficult to tackle than decreasing the feature size (min-cut refers to the minimum number of connecting lines cut in partitioning an IC to reconfigure its layout).

2.2.6 Reliability of Computer Systems

Feature sizes in the submicron range are being etched in 1+ square centimeter dies. Several of these dies are assembled together to form the processing board, the memory board, the hard disk controller, the display image driver, and the printer driver. All of the previous assemblies require high fault tolerance and reliability. If a computer system breaks down, the problem could be at the assembly level or at the ICs manufacturing submicron level. Designing fault tolerant and reliable ICs is a major goal not only because of the costs involved in repairing a faulty IC, but in response to the sensitivity of consumers who are buying a machine the operation of which cannot be seen by the naked eye.

2.3 Magnetic Hard Disks

Storing digital information safely and retrieving it efficiently have been two of the most important attributes of a computer system's storage device. Regardless of how fast the machine processes data, the time the processor spends waiting on data and the accuracy of the information retrieved will always be the most important factors affecting the throughput of the machine and the validity

⁴⁶NP is an acronym for Non-Deterministic Polynomial. It has been conjectured that all NP-complete problems are intractable [25].

of its outputs.

This section elaborates upon magnetic hard disk storage technologies.

- Subsection 2.3.1 defines a storage system within a computer and a computing environment.
- Subsection 2.3.2 describes the magnetic storage technology with a brief presentation of its history.
- Subsection 2.3.3 presents a description of the magnetic hard disk components, their functions and physical structure.
- Subsection 2.3.4 provides a trace of a hard disk data request and an analysis of the factors affecting its performance, in particular the hard disk's response time.
- Subsection 2.3.5 provides some concluding remarks about the magnetic storage technology and about the importance of VLSI as one of the main drivers of the storage technologies.

2.3.1 Computer Storage System

A storage system is considered external to the communication between the processing board and the main memory board, and nonvolatile because it retains the stored data even when the power to the system is switched off. The most widely used nonvolatile storage technologies are magnetic and optical based. With today's state-of-the-art technology, however, magnetic storage has the edge with respect to price, response time, and storage density [101]. It is the storage technology discussed most fully in the following subsections.

The storage system is accessed by an operating system instruction as a result of a user directory read or write command, a process load instruction by the user, or a data-access miss requested by the processor from the main memory during the execution of a certain program. Whether the storage system is a single user or a shared system, each user has a specific storage space within a specified directory of the system. The directory is allocated by the operating system on request by the system's manager or the user⁴⁷. The user can write his/her working files to the allocated directory, read these files, or load programs from it into main memory and execute them.

2.3.2 Magnetic Storage Technology

A magnet is a polarized ferromagnetic material. The polarization has the direction of the magnetizing field, either a north-south (N-S) direction⁴⁸ or south-north (S-N). A permanent magnet retains the polarization even after the magnetizing field is removed and this accounts for the nonvolatility of a magnetic storage system. The digital data bits are stored as polarized magnetic particles, with a 0 bit equivalent to a S-N polarization and a 1 bit equivalent to a N-S one, or vice versa. There are several types of magnetic storage systems, such as magnetic tapes, hard or rigid disk drives, or floppy diskettes. Of the three, the hard disk drives possess the highest data-access rates and best complement a workstation's high speed processor. Hard disk drives also have high linear and areal densities, useful attributes which will be described later.

⁴⁷In certain cases, a directory is allocated on request by a running process.

⁴⁸The north-south direction refers to the orientation of the magnet's poles.

Magnetic Storage History. IBM⁴⁹ pioneered the direct access storage devices (DASDs), in particular the magnetic hard disk technology that we know today [57]. The areal density of DASDs increased by a factor of 100 in the last 20 years [57]. New developments and innovations continue to drive the price per bit of storage down at a rate of 15% to 20%, compounded yearly [3]. Data reliability, performance, and low cost per bit continue to be the main attributes of this technology. Of these three attributes, performance is the most important because it reflects the speed of access to the data stored in the system and dictates the mechanical and media configurations used in the storage device.

Before discussing the factors influencing its performance, a description of the magnetic hard disk basic components setup and operational procedures follows [57, 60, 79].

2.3.3 Magnetic Hard Disk Components

Figure 2.5 shows the enclosure of the storage system containing a stack of rigid disks mounted on a spindle, their corresponding read/write/erase heads with their sliders, and the electromagnetic actuator which controls the movement of the heads over the disks' surfaces. Not shown in the figure is the data channel which feeds data from main memory to disk and vice versa. The air inside the enclosure is filtered so that no particles can interfere between the disk media and the heads during the read/write operations. To control the heat generated during the disks' rotation, an air cooling mechanism is used [57]. The

⁴⁹IBM is an acronym for the International Business Machines corporation.

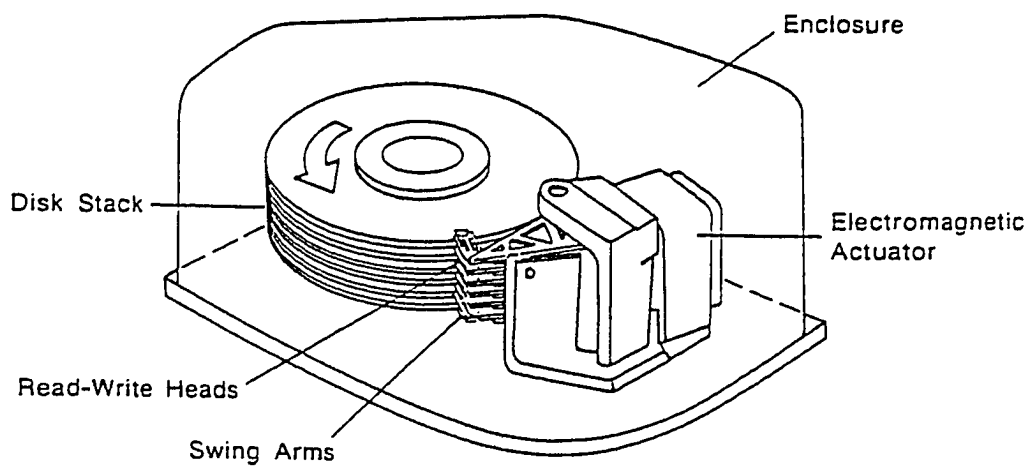


Figure 2.5: Rigid disk file components. Source: [57].

disk substrates are mostly made out of aluminum-magnesium alloys. Glass or ceramic substrates are used for extremely fine head to medium spacings because they provide a smoother surface on which to deposit the magnetic media [57, 79]. Additionally, a thin electroless layer of nickel-phosphorus is plated on the substrates and polished before depositing the media to make the disks surfaces even smoother and flatter. Since the disks rotate at a specified constant speed, their thickness is chosen within the resonance requirements of the disk assembly [57]. Any mild resonance or vibration can induce reading errors and write misregistrations in the system. The coating on the disks⁵⁰ is an alloy of some highly coercive⁵¹ ferromagnetic iron oxide, the layout of which is recognizable by the read/write/erase mechanism of the device. The magnetic layers are deposited on the disks in the form of small, needle-shaped particles or as a magnetic film⁵², and the data stored on them will be retained as long as the disks are not exposed to high heat or fluctuating magnetic fields. To shield the media from air particles and from the heat generated by the disks' rotations, a thin layer of hard carbon overcoat is deposited on top of the media coated disk surfaces [101].

Magnetic Disk Surface Layout. As indicated in Figure 2.6, each magnetic oxide medium is laid in concentric tracks, separated from one another by a constant pitch. An emerging technology uses IC process technologies to etch the tracks on a disk's surface[57] and, consequently, makes possible both finer track

⁵⁰The disks can be coated on both sides.

⁵¹Coercivity is the magnetic field required to reduce the magnetization of a bit to zero.

⁵²Most magnetic films are cobalt-based metal alloys.

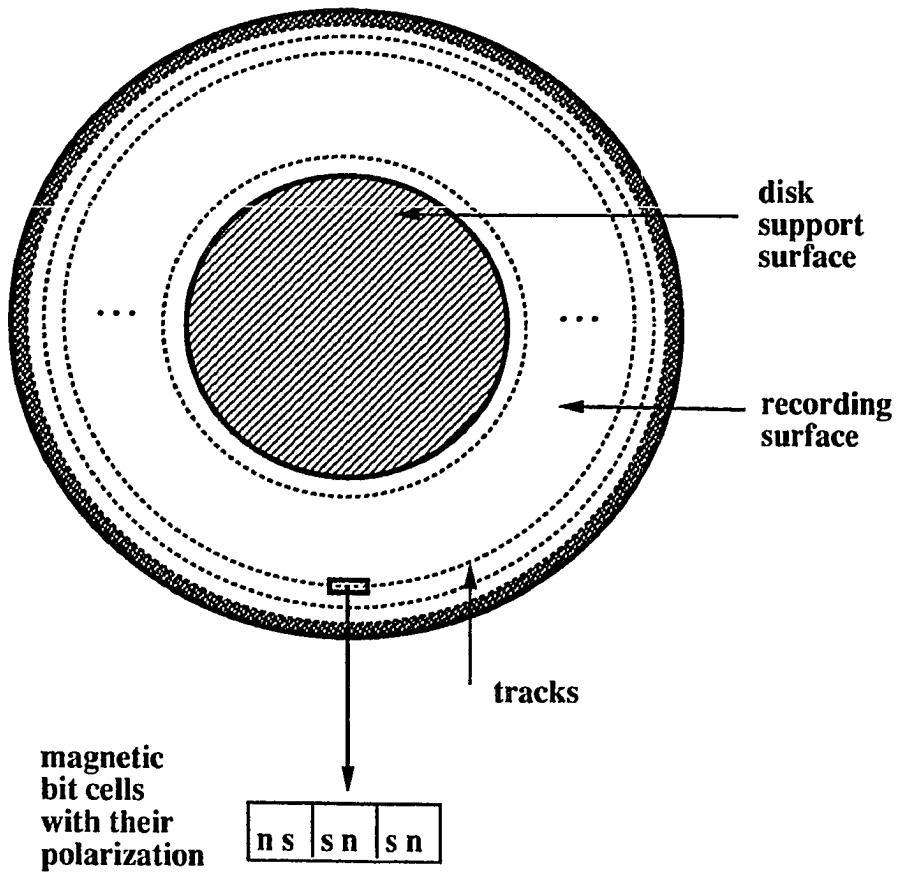


Figure 2.6: Magnetic disk surface layout.

pitches and more clearly defined tracks (a non-homogeneous coating layout could induce reading and writing errors due to the noise generated from the proximity of the bit cells and the tracks [57]). Each track is divided into an equal number of sectors, and each sector⁵³ stores the same number of data bytes⁵⁴ regardless of the sector's radial length⁵⁵. Since the disks are rotating at a constant speed with one head per disk surface, and since the data channels usually have a constant bandwidth, the data access rate is constant and the number of bits per track is constant. The track density, then, is limited to the density of the innermost track. The inner-most track's radius depends on the radius of the spindle support and rotation mechanisms. For a constant data rate, maximum disk areal density is obtained for an innermost track radius equal to half the disk's radius (see Section 2.2 of Mee/Daniel [57] for a detailed proof).

Read/Write Heads. The magnetic heads are attached to slider arms, each dedicated to a disk surface, and supported within a submicron distance from the rotating media by a hydrodynamic air bearing [57]. Most heads are inductive electromagnets and, as illustrated in Figure 2.7, an inductive head is shaped as a ring with a small gap at the surface facing the magnetic medium. The core of the electromagnet is either ferrite or metallic film based. A coil is wrapped around the core, the functionality of which will be described shortly.

Heads having the previous configuration can be used for reading, eras-

⁵³The data on each track is updated one sector at a time.

⁵⁴A byte of magnetic data is a series of eight magnetic cells.

⁵⁵In some disk storage systems where the data channel's bandwidth is not constant, the data recorded on each sector is proportional to the sector's length [33].

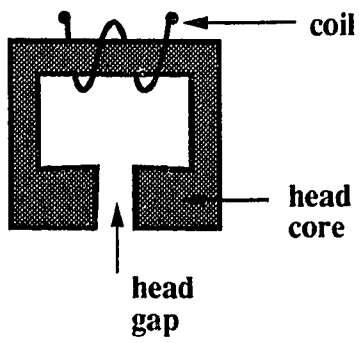


Figure 2.7: Inductive head structure.

ing, and writing data to the media [57]. Ferrite cores are easy to manufacture and have improved wear characteristics. However, at high frequency operations, an all ferrite core prevents a magnetic flux from penetrating it, a problem which some have suggested might be eliminated by laminating it with an insulator like alumina [101]. The resulting head is referred to as the metal-in-gap (M-I-G) ferrite head.

Thin film metal (TFM) based heads have the same configuration as the ferrites, but they are not as easy to manufacture. A photolithography process is used to deposit thin layers of metal on a thick ceramic wafer; at the end of the process, the wafer is cut into small dies which are packaged later as TFM heads [3, 5, 88, 101].

Regardless of the materials used in its construction, the inductive head operates in the same basic way to read from or write to the disk. What follows is a description of the writing and reading processes of inductive heads [57]:

- **The Writing Process**

Since the heads are on a spindle, they can move all together to the corresponding tracks, and a sector⁵⁶ can be written or read on each disk surface simultaneously; but for the purpose of illustration, the writing process of one head on one disk surface will be described. As the disks rotate at a constant speed and the head reaches the corresponding sector on the specified track, a temporary change in the coil⁵⁷ current induces a magnetic

⁵⁶As mentioned earlier, writing and reading from a rigid disk is performed one sector at a time and not one byte or several bytes at a time; a sector is a multitude of bytes, the number of which is dependent on the storage system.

⁵⁷The coil is wrapped around the head's core.

field in the head's core. The direction of the magnetic field in the core depends on the direction of the current in the coil; one current direction corresponds to a 0 bit being registered and the other a 1. With a specified direction, the magnetic field attempts to cross the core's gap, polarizes the magnetic bit cell passing under the head, and a bit is written⁵⁸.

- **The Reading Process**

As the disks rotate at a constant speed and the head reaches the corresponding sector of the specified track, a magnetized bit cell, passing under the head's gap, induces a magnetic flux in the core. In turn, the changing core flux induces a certain voltage in the coil. The value of the voltage is interpreted by the disk's circuitry and is dependent on the magnetic bit cell read. For a more detailed look at different read and write mapping techniques on magnetic hard disk media, consult Section 2.2 of Mee/Daniel [57].

Both, the M-I-G and the TFM heads are suitable for high data rates and high bit density systems, and the fact that they are not touching the rotating disks adds to the reliability of the stored data and to the durability of the media. The number of heads per storage system depends on how many disks the system has, whether each side of the disks is magnetically coated, and whether each head can perform read, erase, and write functions. A head per track configuration was suggested at one time in the past, but dropped later because its implementation was not cost competitive with the DRAMs [57]. For

⁵⁸The erasure of the bit is overridden by the writing process.

extremely high bit densities and high data rates disk systems, separate read and erase/write heads were suggested, each with its optimized functional capabilities. Magnetoresistive read-only and inductive write-only heads have been developed for these purposes and are currently implemented in IBM's top-of-the-line hard disk storage systems [101]. For more details on the operations of the dedicated heads, refer to [3, 5, 57, 88, 101].

Actuator and Servomechanism. The heads' movement over the rotating disks is a mechanical action, controlled by a head-positioning servomechanism. The servo determines the distance the heads must travel until they reach the right tracks, initiates the heads' movement, and stops the heads when the specified tracks are reached. In most rigid disk systems the location of the data on the disks is recorded on one disk, called the servo disk [57]. When a read or write command reaches the servo, the location of the data to be read or written is determined from the servo disk, and an electromagnetic actuator is instructed to move the corresponding disk heads to the specified track coordinates. Also, the actuator must keep the heads on track to minimize misregistration or misreading errors [57]. A discussion of the actuator's design and the servomechanism's circuitry is in Section 2.2 of Mee/Daniel [57].

2.3.4 Magnetic Hard Disk Performance

The performance of a magnetic hard disk is a measure of how fast it comes back with the requested data. The response time of the hard disk has embedded in it all the attributes of the mechanical, electrical, and magnetic properties of the storage system. What follows is a trace through a data request

and an evaluation of the factors that affect the storage system's response time.

Stepping through a Data Request. The hard disk has a FIFO⁵⁹ queue for incoming data requests. The top request in the queue is accessed by the disk's controller, which translates it into a servomechanism command. The servomechanism finds the corresponding sector by searching for it in the servo disk⁶⁰. Once the corresponding disk and the sector location on it are found, the servomechanism computes the corresponding distance the disk surface's head has to travel from its current position. The travel distance of the head is then passed to the electromagnetic actuator, which controls the head movement until it reaches the corresponding track⁶¹. Once the target track is reached, the head has to wait for the corresponding data sector to come under it⁶²; the head may miss the sector. If so, it might have to sit and wait one disk revolution before it reaches the specified sector a second time. Once the sector is found, the data transfer occurs, either a data read or a data write⁶³.

Magnetic Hard Disk Response Time. After tracing a hard disk request procedure, the response time of a magnetic hard disk storage system is expressed as follows:

$$T_{response} = T_{service} + T_{wait} \text{ (sec)} \quad (2.4)$$

⁵⁹FIFO is equivalent to the First-In First-Out paradigm.

⁶⁰The time to termination of this procedure is referred to as the actuator positioning time.

⁶¹The time to termination of this procedure is referred to as seek time.

⁶²The time to termination of this procedure is referred to as latency.

⁶³The time to termination of this procedure is referred to as the data transfer time.

where the wait time is the time the request had to wait in the queue. The service time's expression is:

$$T_{service} = T_{actuator} + T_{seek} + T_{latency} + T_{miss} + T_{data-transfer} \text{ (sec)}. \quad (2.5)$$

The service time in Equation 2.5 has been improved over the years by innovations and enhancements to the mechanical, electrical, and magnetic components of the storage system, for instance:

- The data transfer and the actuator positioning times have been improved by using a harmonious combination of mechanics and electronics, called micromechatronics, to put together the head, the rotational motor, and the control servomechanism into one single device [88].
- The magnetic bit cell length has been decreased and with it the gap width of the head's core decreased, the linear densities of the media⁶⁴ increased, and the data transfer time decreased. A new bit cell layout technique aligns the cells perpendicularly instead of longitudinally along the track, which increases the bit cell densities (this layout technique was not possible in the past due to unavailability of media that could handle such a magnetization [57]). Unfortunately, the increase in the bit cell densities caused bit interference noise and bit misregistration and misread problems during hard disks accesses. To preserve the storage system's data reliability, smaller head designs with very small core gaps were implemented [49, 57, 101]. The noise problem was handled by decreasing the head medium spacing [57, 101].

⁶⁴Linear density is expressed as the number of bits per millimeter.

- The track density⁶⁵ has been increased by improving the head technology and the layout process of the tracks. Magnetic film media and ICs etching processes have been used to obtain a finer track layout and a smaller track pitch [49, 57]. The increase in track densities induced an increase in the areal density of the disk—equal to the product of the linear bit density and the track density—and a decrease in the seek time. Unfortunately again, the increase in track densities caused track to track noise interferences during a write or a read operation, problems that were solved by better actuator and servomechanism designs and by the development of smaller heads with very small core gaps [49, 57, 101].
- Multiple track access implementations, coupled with faster disks rotational speeds, have steadily decreased the data transfer time [49, 57]. With the increase of the data transfer rate⁶⁶, the channel circuitry has improved and faster data processing ICs have been employed to handle the higher bandwidth between the storage system and the main memory.

2.3.5 Concluding Remarks

Until very recently, paper was the dominant form of storage [49]. The digital information storage technologies have, nonetheless, become cheaper, more reliable, more easily accessible, and more user friendly, encouraging business, government, and private sectors to rely upon them more and more.

⁶⁵The track density is expressed as the number of tracks per millimeter.

⁶⁶The data transfer rate is expressed as the number of megabytes per second.

Note once again, however, that VLSI technologies are integral parts in the head manufacturing process and the disk control circuitry. Lithography is being used in the manufacture of thin film heads and digital signal processors, and caches are being used in disk controllers to reduce the number of wait-on-data cycles of CPUs [105]. Since DRAM speeds have not kept pace with microprocessors, further improvements to the storage attributes might help to compensate for lag in DRAM performance improvements, allowing the overall system performance to keep increasing.

2.4 Color CRT Displays

With the introduction of multitasking⁶⁷ capabilities into the workstation environment, the need for larger, faster, and appealing computer terminals emerged. The display technology best positioned to take the lead in the merger of computer and display hardware was cathode-ray tube (CRT) based. What follows is a brief presentation of the CRT's history (Subsection 2.4.1), a description of other display competing technologies (Subsection 2.4.2), a presentation of the color CRT components (Subsection 2.4.3), an illustration of the generation techniques of a screen image and its elements (Subsection 2.4.4), a presentation of the factors affecting the CRT's bandwidth (Subsection 2.4.5), a description of the CRT's graphics adapter components (Subsection 2.4.6), and, finally, a presentation of the CRT's resolution and the technical barriers facing the CRT industry as it attempts to achieve maximum resolution (Subsection 2.4.7).

⁶⁷In multitasking, a single user can interact with several computer applications concurrently.

2.4.1 CRT History

The CRT's invention in 1879 is credited to William Crookes [90]. More than a century after its invention, CRT technology still has the highest market share in the display industry [91], receives the highest research and development (R&D) expenditures, and demonstrates the greatest improvements per R&D dollar spent [23]. CRT based displays replaced the typewriter and the teletype terminal displays of the 1960s because they facilitated user interaction with time-shared computers⁶⁸, and had the fastest response times [20]. Several computer aided design (CAD), animation, simulation, and layout applications were developed in the late 1970s and only the CRT display provided the color imagery necessary to capture the illusion of reality in animations, simulations, and designs. Its large screen size enhanced the multitasking attributes of the machine⁶⁹ and delivered the fast response time that provided for an efficient interaction between the user and the computer [55].

CRTs are available in black and white (B&W) and in color image display capabilities. The first color CRT was produced in 1950 [90], and had the B&W CRT technology as its backbone. Today, the CRT still holds the highest market share among the total display technology market—71% as of 1989—and it is not projected that other technologies will challenge its reign until the mid-1990s [91].

⁶⁸Time-sharing is an operating system function which schedules the processing time of the computer among several users, each with a particular computer time slot.

⁶⁹Controlled by a window manager, several user applications can be launched, each in its separate window, and with all the windows showing on the large screen.

2.4.2 Competing Display Technologies

Several display technologies competed with the CRT technology for market share in the past, but it was not until the age of the personal computer that the market witnessed a flood of innovations from display companies around the world. Of the display technologies competing with the CRT's, a few show some promise, and Japan is the leader in their development and market share [23]. The most promising of these technologies is liquid crystal (LC) based. LC displays (LCDs), available in black and white and in color, are compact and low in power consumption, are increasing in size and decreasing in cost, and their display elements are easily addressable with thin film transistors. But improvements are needed to increase their resolution and to enhance the speed of their picture updates for animation [23, 44, 78]. The other two potential technologies are electroluminescent displays (if full color ones are developed) [87] and plasma displays (once their technology matures and affordable color capabilities comparable to the CRT's are achieved) [23]. For a more detailed discussion of these competing technologies, see [23, 44, 78, 87, 90, 91].

2.4.3 Color CRT Components

As shown in Figure 2.8, a CRT consists of several parts [75, 90]: the bulb's⁷⁰ faceplate or viewing screen upon which phosphor dots are deposited, the funnel, and the neck. In addition, in color CRTs, a metal shadow mask is suspended behind the faceplate, between the electron beams and the color

⁷⁰The bulb is also referred to as the bottle or the envelope.

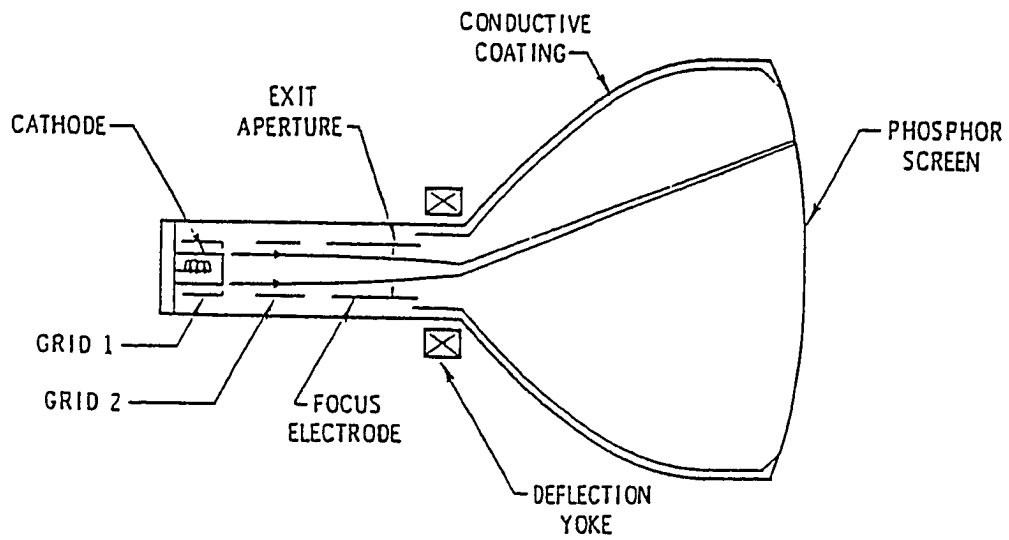


Figure 2.8: CRT display components. Source: [90].

phosphor dots. The metal shadow mask, not shown in Figure 2.8, acts as the coordinate system for the image displayed on the screen of the CRT. What follows is a description of the previously mentioned components and their role in producing the image displayed on the CRT's screen.

Bulb's Faceplate. The bulb's faceplate defines the surface of the viewing screen and isolates the inner CRT components by maintaining the vacuum needed for the cathode rays to hit the inner side of the screen without any particle interferences. Normally, the vacuum filled bulb is made of glass. However, in certain cases, the funnel and/or the neck may be made of ceramic or metal, depending on the CRT application [90]. In most cases, the screen is designed such that a 4:3 aspect ratio is satisfied, where the horizontal side of the screen measures $4/3$ times the vertical side. The size of the screen, usually measured as the screen diagonal, is limited by the ability of the vacuum filled glass to withstand the atmospheric pressure on its faceplate. The vacuum in the bulb is maintained at a specified level by an antenna getter⁷¹ [54].

Bulb's Neck and Funnel. The second major part of a color CRT is the neck of the bulb. The neck area houses three cathode heaters, three cathodes, three electrostatic beam⁷² accelerating apertures, and an electrostatic beam focusing aperture. In certain cases, the beam focusing aperture is magnetic and positioned outside the tube's neck. Three electron beams are needed to excite three

⁷¹A getter is defined in SDEET as a substance introduced in an electron tube to increase the degree of vacuum by chemical or physical action on the residual gases.

⁷²The beam referred to is the ray of electrons generated by the electrode.

types of phosphors, red, green, and blue, simultaneously, and create the corresponding color on the viewing screen by an additive effect [90], a process that will be described later.

A cathode ray is generated in the vacuum by heating a barium compound, or a barium oxide **cathode**⁷³, to temperatures that enable the release of a high density **electron beam**, and by applying an electric field to the hot surface of that cathode. The electric field is generated by applying a positive voltage or potential to the anode of the **accelerating aperture**. The accelerating aperture has two parts: the first is the accelerating anode, which generates the positive potential to draw the electrons from the cathode surface, and the second part is the negative potential electrode, which controls the intensity of the electron beam. The negative potential electrode is sandwiched between the cathode and the accelerator anode and controls the beam intensity by applying a negative potential electric field on the surface of the hot cathode; i.e., as the negative potential increases, three consequences ensue: first, the repelling forces of the negative field on the cathode increase; second, the anode's positive field effect becomes weaker; and third, the intensity of the electron beam decreases. (Chapter 6 of Tannas [90] contains a discussion of several other accelerator structures.) The intensity of the beam affects the luminance and the resolution of the CRT, effects which will be described later.

While still in the bulb's neck, the electron beam leaves the acceleration aperture and goes through a **focus aperture**. The focus aperture can be

⁷³A cathode is an electrode at which negative charges are formed.

magnetic⁷⁴ or electrostatic, but its task is to orient the beam toward the screen by overcoming the electrons scattering⁷⁵ and drifting⁷⁶ effects. Several focusing paradigms are discussed in detail in Chapter 6 of Tannas [90]. While leaving the bulb's neck, the electron beam passes through a **deflection aperture**. The deflection can be induced magnetically or electrostatically, and it is applied as vertical and horizontal fields on the beam to position it to the desired coordinates on the viewing screen. The viewing screen is maintained at a high positive potential and acts like an anode for the drifting electrons through the **funnel** of the bulb, which constitutes the third CRT component.

Screen Phosphors and Electron Beams. The geometrical positioning of the three electron beams with their corresponding generators, accelerators, and focus apertures depends on the color phosphor dots layout on the screen. These phosphors constitute the fourth major component of the color CRT. The array of colors on the screen is generated by adding three color sources, red, green, and blue (RGB). The phosphor RGB color sources pass through two stages to establish the cathodoluminescence phenomenon:

1st Stage. As soon as the deflected electron beams drift into the funnel space (shielded from the earth's magnetic field by an internal canceling magnetic field) and hit the corresponding phosphor coordinates on the screen, the phosphors are excited by the energy with which the electrons are bom-

⁷⁴Magnetic focus apertures have had the highest resolution performance results.

⁷⁵Scattering occurs due to the electrons or the similar charges repulsion effect.

⁷⁶Drifting occurs due to the positive potential field while passing through the accelerator anode.

barding them and emit a fluorescent radiation.

2nd Stage. Soon after the excitation ceases, the phosphors emit a phosphorescent radiation, completing the cathodoluminescence phenomenon.

Each of the three electron beams is dedicated to phosphors of a particular color, and all the phosphors on the screen have to go through that excitation process for an image to be created. The colors generated depend on the beams' intensity which can be controlled by the accelerating apertures of the CRT. As the beams drift in the funnel space, though, they experience scattering effects due to the electron repulsion phenomenon, which brings about the need for a fixed and defined phosphor coordinates scheme.

Metal Shadow Mask. The screen coordinates system takes the form of a metal shadow mask perforated with holes or slots, as shown in Figure 2.9. This mask is placed right before the phosphor covered screen and constitutes the fifth major CRT component. The perforated shadow mask works as follows⁷⁷. Each hole in the mask covers an RGB triad of phosphors. The triads in front of the mask are arranged uniformly by a photoresist process. The photoresist process enables the layering of phosphor dots, one color at a time, where the photoresist covers the previous layer/s of dots while one of the remaining layers is deposited. The latest phosphor depositing process technology uses thin-film phosphors because of their homogeneous surface and the image enhancements they provide to the CRT [80, 90]. Thin-film phosphors are mostly inorganic

⁷⁷A slotted mask configuration has been used in the Trinitron CRT technology developed by Sony Corporation. It is not described here.

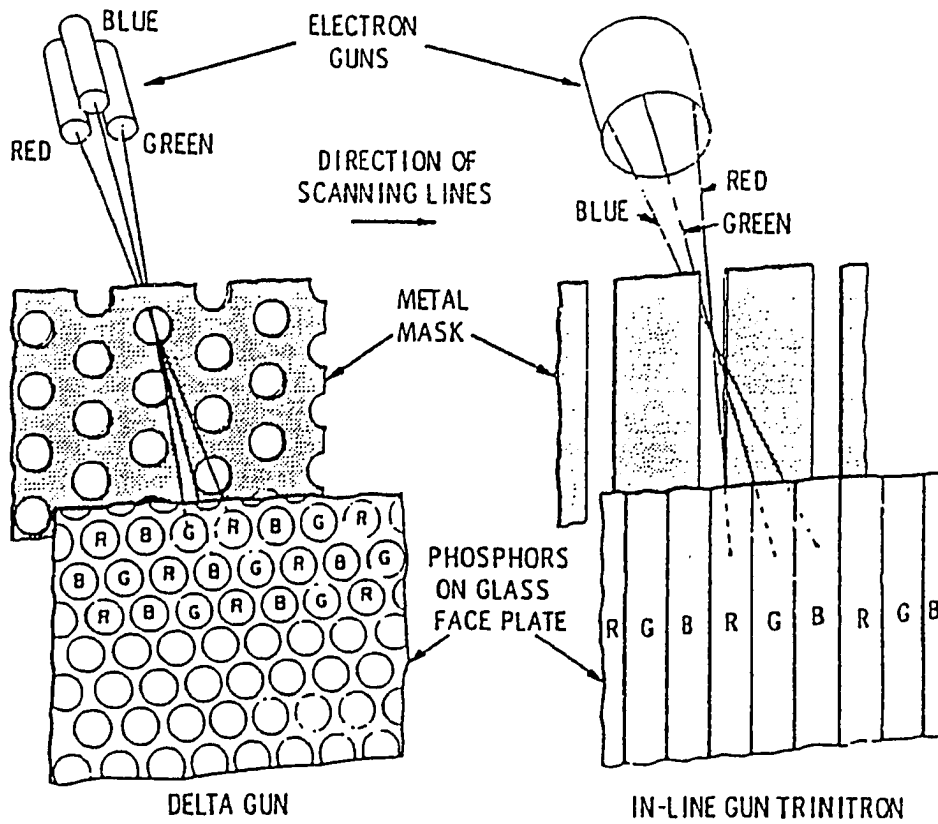


Figure 2.9: Inner screen phosphors layout structures. Source: [90].

crystalline materials, with a few added impurities to enhance their activation efficiency during the electron excitation process (refer to Chapter 6 of Tannas for details on the different types of phosphors and their attributes [90]). After applying the phosphor film on an optically smooth faceplate, a thin coating of aluminum is deposited on the film to enhance its conduction [90].

Figure 2.9 shows the electron guns positioned in a delta configuration to excite simultaneously the triangularly shaped RGB phosphor triads on the screen. As soon as the beams hit the phosphors, each with a particular intensity, the additive light emitted from the phosphors forms the final transmitted color of the image's dot. The limit on the number of phosphor dots that can be excited on the screen is a function of the metal mask's hole pitch. The smaller the hole pitch, the harder it is to manufacture the mask⁷⁸ and the more expensive the CRT becomes. By providing the mask, the definition of the image displayed is a function of both how well the circuitry of the CRT can focus the beam to hit the screen dots through the mask holes and the total number of holes drilled in the mask.

2.4.4 Screen Image

The beams can scan all the holes on the screen in two ways. The following paragraphs illustrate the two image generation techniques and describes the basic image element, the pixel.

⁷⁸Smaller hole pitches induce lower yields of good masks.

Image Generation Techniques. The first image generating technique is called interlaced scanning, where every other horizontal line of holes is scanned in the first pass of the beams and the nonscanned lines are scanned in the second pass. A line is equivalent to the number of holes perforated in the mask horizontally. Non-interlaced scanning is the second image generating technique used, and, in this scheme, no lines are skipped while scanning the screen. The electron beams scan the screen in one fixed direction; after scanning a line, the electron beams go into a blanking stage before they start scanning the next line. Due to flicker and phosphor light emission persistency limitations, each image on the screen has to be scanned, or refreshed, several times a second, depending on the CRT cathodoluminescence parameters. When interlacing is used, an image is scanned half as many times as the frequency of each pass (i.e., if each pass is performed eighty times, with a two-pass full image scan, the whole screen is effectively refreshed forty times). Most CRTs today use non-interlaced scanning because it provides sharper images and the available screen-driving hardware is capable of performing it, and it is cheap enough to be incorporated as a basic component of the workstation system [75, 90].

The Pixel. Each image is composed of picture elements, referred to as pixels. It is important to note that the image is in the computer memory and is being fed to the screen by the CRT's hardware driving circuitry. Each pixel is mapped onto the computer memory as a series of bits. The bits are transformed into analog signals that control the intensity of the electron beams and, in turn, the color of the pixel.

2.4.5 CRT Bandwidth

The rate at which the pixels' bits are fed to the controls of the electron guns determines the rate by which the screen can be scanned, or the bandwidth. The bandwidth is equivalent to the rate at which the electron guns can be switched on and off. Special attention must be given to high switching electron guns rates due to the overheating that occurs in the deflection apertures [59]. Several cooling techniques have been employed in the more sophisticated CRTs [59, 90]. The guns are assumed to switch on and off every time a pixel is scanned, because different beam intensities are supposed to be generated.

The maximum bandwidth is equal to the number of holes in the shadow mask multiplied times the refresh rate of the screen. The refresh rate of the screen is the number of times the screen is scanned per second to avoid flicker. The maximum bandwidth is expressed in Equation 2.6 as:

$$BW_{max} = (HS_{in} * HPI) * (VS_{in} * HPI) * RR_{Hz} \text{ (Hz)} \quad (2.6)$$

where HS and VS are the horizontal size and vertical size in inches of the CRT screen, HPI is the number of mask holes per inch, and RR is the refresh rate in Hertz. The number of holes in the metal shadow mask, or the maximum number of pixels that can be displayed, is equal to the number of holes in a horizontal line multiplied by the number of holes in a vertical line.

2.4.6 CRT Hardware Drivers

Hardware drivers of a color CRT are components of the graphic subsystem of the computer workstation, and they are typically assembled on a circuit

board placed in the workstation cabinet or in the CRT assembly itself [9, 67, 102]. They are made up of four components. The first component of the hardware drivers is the **controller**, which handles the interface between the computer and the peripherals, such as the keyboard or the mouse, and computes the coordinates of the pixels to be echoed later as the scanned image on the screen. The second component is the pixel memory storage, or **buffer**, which is constantly updated and accessed in the refresh cycles of the screen. The availability of very fast DRAMs allowed CRTs to replace the display technologies of the 1960s like teletypes and typewriter terminal displays [55]. The third component is the **video generation system**, responsible for feeding the pixels' serial bit stream to the D/A⁷⁹ converter to generate the RGB color analog signals, for refreshing the screen, and for synchronizing the CRT's response to the analog signals just created, in particular the electron guns' intensities. The fourth component is a **supervisory processor** which coordinates the timing and the execution of the tasks previously mentioned. Several hardware drivers can be designed for specific applications of the CRT, and these drivers could become a major cost overhead to the total cost of the workstation [33]. A recent enhancement of the hardware driver's attributes has been achieved through the use of cache memory to accelerate the image refresh rates and timing and to improve the resolution of the display as a whole.

⁷⁹A D/A is a digital to analog converter.

2.4.7 CRT Resolution

Resolution has become the key parameter in any display's technical description, in particular CRT based displays. Estimated to double every five years in the future [76], resolution captures the performance characteristics of all the CRT's major components and the hardware driving it [96]. The Standards and Definitions Committee of the Society for Information Display defines resolution as a measure of the display's "ability to delineate picture detail." In other words, it is a measure of the number of pixels displayed per image [90].

To increase the CRT's resolution, several factors and technical barriers must be taken into consideration including: 1) the hole pitch of the metal shadow mask, 2) the switching speeds of the electron guns, 3) the expansion of the metal shadow mask, and 4) the misconvergence of the electron beams.

Metal Shadow Mask Hole Pitch. The maximum attainable resolution is equal to the number of perforated holes in the metal shadow mask. So, for a fixed screen size CRT, increasing the maximal resolution is synonymous with decreasing the mask's hole pitch. But the smaller the hole pitch, the harder it becomes to manufacture the mask. The maximum resolution per inch that a human eye can discern is equal to 300 holes per inch [40], or a hole pitch⁸⁰ of 0.085 millimeter. For a color CRT with a 19-inch diagonal and a 4:3 aspect ratio the maximum resolution requires for nearly 15.6 million holes to be drilled in the mask, or approximately 4000 holes on the side for a 1:1 aspect ratio! By today's standards, drilling such a huge number of tiny holes in the metal shadow

⁸⁰The shadow mask hole pitch as of 1991 is 0.28 millimeter.

mask is extremely difficult. The yield of good and non-defective masks at such a high resolution is so low that the manufacturing costs of the corresponding color CRTs are prohibitively high.

Higher resolutions might be economically feasible for very small color CRTs [39], or B&W CRTs where no shadow mask is required [76]. But for 14- or 19-inch workstation displays, several technical problems other than mask yield remain.

Electron Guns Switching Speeds. The second technical barrier to increased CRT resolution is switching the electron guns quickly enough so that no flicker occurs at such high resolutions [40, 76]. Given the 19-inch color CRT screen with 300 holes per inch resolution, the maximum bandwidth of the guns is close to 1.56 gigaHertz for a screen refresh rate of 100 Hertz. Today's multi-guns structure and design are not adequate to handle these switching speeds, and several material reformulations are needed to handle the amount of heat generated from the screen scanning process [13, 40]. Primarily, the phosphors that are used need enhancements. At such a high bandwidth, the phosphor dots are being excited for a very short period of time, so only phosphors with smaller electric current⁸¹ requirements will satisfy the maximum resolution limitations.

Metal Shadow Mask Expansion. Developing metal shadow masks that will not expand as their temperature increases is also problematic in increasing CRT resolution. As the electrons bombard the screen through the mask, heat is

⁸¹The electric current is equivalent to the electron beam.

generated and induces metal expansion [13].

Electron Beams Misconvergence. The fourth obstacle to maximizing resolution for workstation size displays is the electron beams' misconvergence [65]. Since the phosphor dots are smaller, the separation of the electron beams has to be increased to focus properly on the dots. However, this larger separation of the guns creates beam misconvergences on the screen.

Depending on the CRT's application, tradeoffs have to be considered in the final design to determine the degree of focus sacrificed for greater convergence of the beams. By today's standard designs, cathode materials will not pose a problem in attaining such high switching speeds [76]. The hardware drivers will not be an obstacle either, because VLSI's evolution brings with it faster microcontrollers, faster and cheaper DRAMs, and more efficient paradigms to refresh the screen at high rates for a large number of pixels [8, 14].

2.4.8 Concluding Remarks

The CRT belongs to the class of active, direct-view refresh displays. Its performance and cost are affected by the type of metal and hole pitch of the shadow mask, by the screen phosphors' brightness⁸², persistency, and deposition technique, by the hardware driving circuitry and how well it matches the resolution and bandwidth parameters of the CRT, and by the overall design of the bulb [13].

⁸²Brightness describes the perpetual effect generated by the luminance and the chrominance of the light emitting object.

The disadvantages of a color CRT are its high power consumption, its curved face plate, its volume, and its weight. Over the years, its high switching speed, its high resolution, and its low cost have kept its market share high. It will remain the technology of choice for many workstation applications, especially in CAD and simulation markets. Currently, Japan is the leader in enhancing the CRT attributes [23] and, from the R&D money already invested in it, the CRT will most likely retain a high market share through the 1990s [16].

2.5 UNIX Operating System

The 1990s wave of information processing is fueled not only by advances in hardware capabilities—in the form of faster processing speeds, faster and denser memories, and high bandwidth communication links—but also by open system oriented software. The ability to develop open system software has been the key to creating new waves in the industry because it promotes the notion of standardizing the interface between different computer systems and the interface between each system's components [17]. The ultimate goal of open system programming is to use one software with all computer systems without incurring further development or porting costs. Development-from-scratch costs are incurred when the whole software for a particular system is rewritten, usually at the request of a company that manufactures the system⁸³, or when some software functions are added to the originally developed version to enhance its user demand. Porting costs are incurred when parts of the software that handle

⁸³For example, IBM developed AIX, a version of the operating system UNIX, to be used in its RISC System/6000 workstations.

its interface with the hardware are rewritten to adapt it to the configuration of the host computer system. The previous software issues are most relevant in the design of computer operating systems (OSs). A brief history and design structure of the most successful workstation environment and open system software, UNIX, is presented below [17, 70]. The description of the design structure includes a trace through a UNIX shell command and a presentation of the most common UNIX managed functions.

2.5.1 UNIX History

UNIX has become the standard workstation OS component and the first open system software to receive wide acclaim from the scientific and programming communities [17]. Developed at Bell Laboratories by Ken Thompson and Dennis Ritchie in the late 1960s [70], UNIX did not gain popularity and market share until, in the 1970s, DARPA⁸⁴ funded the University of California at Berkeley to develop a standard UNIX system for government use. The Berkeley UNIX system was supposed to provide networking support for DARPA's Arpanet and local area networks (LANs). Since the Berkeley UNIX system was a government funded unclassified project, the public had access to it as free software. The workstation companies, in particular Sun Microsystems, Inc., capitalized on that opportunity and used the Berkeley UNIX in their first workstations. Aside from being free, UNIX was chosen as the workstation's OS of choice because it provided the software support for networking, multitasking, and distributed computing, had a small size, and was laid out in a modular,

⁸⁴DARPA stands for Defense Advanced Research Projects Agency.

clean design [4, 29, 70].

2.5.2 UNIX Design Structure

As a computer resources and network manager, Berkeley UNIX was designed in two parts, the system programs and the kernel [70].

System Programs. Figure 2.10 illustrates the system programs which constitute the first layer of the OS and handle the interface between the user and the kernel. The user interacts with the computer through this layer by creating an environment where he or she can issue commands to the hardware to execute and receive a response. This environment in UNIX is called a shell. A user can create many shells, and, with the advances in computer/user interface software, several “window”-based interactive software applications have been developed wherein each window is a shell in itself.

As soon as the user issues a command, it is interpreted and matched within a library of commands. If the library validates the system’s acceptance of such a command, a system call is issued to the kernel to access the command’s compiled code and to execute it while being assisted by the computer’s hardware. UNIX system calls are categorized as process⁸⁵ control calls and file and information manipulation calls [70]. Most of the system programs in the shell commands library are written in C, which facilitates moving the software to different hardware platforms and eliminating the hassle of binary code com-

⁸⁵In [70], a process is defined as a program in execution.

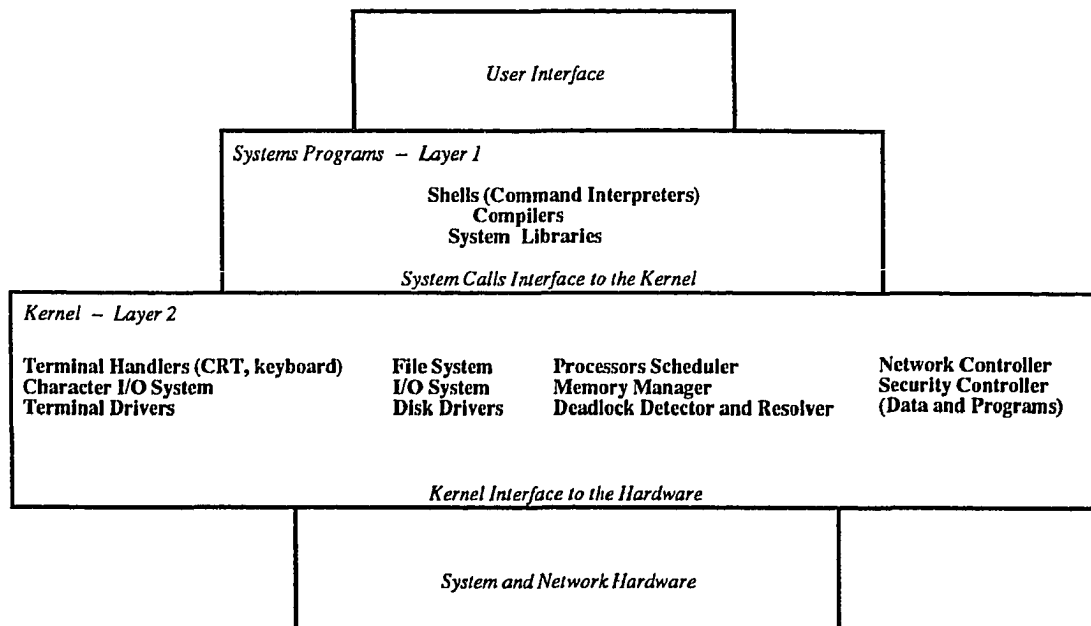


Figure 2.10: UNIX layer structure.

patibility and portability. If the OS is moved, all that is needed is a C compiler on the host machine to map the C source code of the programs to object binary codes understandable by the machine. (Depending on the hardware platform some software porting might be required.) Source code is usually a synonym for a program written in a high level language. Object code is usually a synonym for a machine assembly code of a program.

Kernel. After being interpreted by the first layer, system programs are passed as system calls to the kernel, or the second layer, which handles the interface with the computer's hardware. OS porting is usually performed at this level, if needed. As shown in Figure 2.10, many tasks are performed by the kernel, the first of which is managing the I/O of the computer, in particular, between the processors, the main memory, the hard disk, the monitor, the printer, the mouse, and the keyboard. Other forms of I/O occur at the local or wide area network level. Before discussing the networking capabilities of UNIX, a brief description of the execution of a command in a UNIX managed system is in order.

2.5.3 Stepping through a UNIX Command

Suppose that the command requests the user's current directory content to be listed (the *ls* command). The keyboard is used as the standard input device, and the monitor as the standard output device. As the command is typed, the shell feeds it to the screen buffers, and the characters of the command are echoed on the monitor. As soon as the carriage return is pressed, the command is interpreted and forwarded to the kernel. The kernel queues the

request in the file system, usually a magnetic hard disk, with all the coordinates of the corresponding directory⁸⁶ and waits. The disk returns the requested data to the kernel in a buffer. The kernel, correspondingly, transfers the data to the main memory. In turn, main memory feeds it to the screen update buffers, and the directory listing is scanned on the monitor. There are several configurations for establishing the link between the kernel and the I/O devices, a detailed presentation of which can be found in Peterson/Silberschatz [70].

2.5.4 Main UNIX Managed Functions

The main functions handled by UNIX are memory management, multitasking, detecting and resolving deadlocks, security, network communication, and distributed computing.

Memory Management. When executing shell commands, the management of main memory and permanent storage memory is handled by the operating system, especially during program execution. When the data requested by the program is not available in main memory, the OS issues the hardware calls to fetch it, and make it available to the running program as efficiently as possible; again, consult [70] for more details.

Multitasking. While waiting for data, the kernel facilitates the multitasking operation of the workstation by efficiently scheduling the processor/s for different

⁸⁶The directory system in UNIX follows a tree structure, with several tools to protect it and augment it.

user requested tasks. Multitasking is performed when a single user is running more than one application in one or several shells, and the OS has to swap the applications' processes in and out of main memory for the processor/s to complete or update their execution. For each of these applications, the processor must be scheduled for a certain period of time, depending on the priority of the application, its size, and the main memory it requires.

Deadlocks Handling and Security. Other computer managerial issues face the OS, the most important of which are detecting and resolving deadlocks⁸⁷ and protecting the overall computer system from being infiltrated by computer pirates. Especially with the rapid growth of computer networks, UNIX-provided security has become a priority for privacy and data protection requirements.

Network Communication and Distributed Computing. UNIX facilitates the setup of local area networks, or workstation distributed computing environments, because it has the kernel communication routines which enable the launching of different processes⁸⁸ on different workstations, with each process reporting to the same originating machine. Connected on a Ethernet or a token ring, the workstations can communicate with each other over the network's fiber optic lines and with other networks via communication protocol routines performed by the kernel of each network. Usually a network has a file server which

⁸⁷A deadlock occurs when a process, called P1, requests a hardware resource in order to terminate, and that resource is held up by a process, called P2, and P2 requires P1 to finish in order to terminate.

⁸⁸In the literature, these processes are referred to as remote procedure calls or remote login sessions.

acts as the second OS of all network connected machines and as their gateway to the rest of the world. (For more details regarding distributed computing and its issues, see Bertsekas/Tsitsiklis [6] and Peterson/Silberschatz [70].)

2.5.5 Concluding Remarks

Since the American Telegraph and Telephone corporation (AT&T) developed the first version of UNIX in the late 1960s, several other compatible versions from AT&T and other companies have emerged. DEC developed its own Ultrix OS for its DECstations; IBM developed AIX for the RISC System/6000 machines; Sun Microsystems, Inc. developed SunOS for its SUN and SPARC series of workstations; Microsoft developed Xenix for PCs; and several other organizations like Uniforum and Usenix have helped in marketing the UNIX OS to different hardware platforms other than the workstation's. However, the diversification of UNIX development houses created instability in a market that is searching for a standard, off-the-shelf version of the operating system [32]. To deal with the diversity, Posix was developed as a platform of technical standards for users and programmers, and it provided the features and the characteristics the UNIX operating system should have to become universally accepted, across all available hardware platforms [32].

Chapter 3

Workstation Supply Models

This chapter describes models for the supply of workstation components and workstation assembly. Section 3.1 reviews the simulation approach used. Section 3.2 presents the discrete event simulation supply models of microprocessors and DRAMs, Section 3.3 magnetic hard disks, Section 3.4 color CRT displays, Section 3.5 UNIX operating systems, and Section 3.6 the linear workstation assembly process model. Several components attributes and trends will be presented, and the definitions, terminology, and concepts detailed in Chapter 2 will be referred to as this chapter progresses.

3.1 Simulation Overview

Before presenting the supply models, a discussion of the simulation modeling approach adopted in this dissertation is presented in Subsection 3.1.1, and a diagrammatic approach to communicating the model's structure is given in Subsection 3.1.2.

3.1.1 The Simulation Modeling Approach

Implementing an idea or a system design involves labor and material costs and, if the system undergoes several design or structural changes, these costs can be incurred several times over. Simulation is a tool to model the

behavior of the system, experiment with it, and tune it to obtain the desired behavior without incurring the material costs. Certainly there are labor costs incurred in developing the simulator, but these costs are offset by the reduction of risk and unpredictability associated with the behavior of any new design or any modification to an old design [35, 56]. Once the system's inputs, outputs, and its inner working characteristics have been defined, the effects of any change to the system can be traced through the mathematical equations in a way understandable by the simulation tool, and the consequences of that change can be analyzed.

Computer simulation, by its nature, involves a certain amount of art in creating the model. Once created, though, the model can be systematically used to increase our understanding of the actual system's dynamics through simulation experiments. In itself, simulation is not an optimization tool [56], but several simulation techniques can be incorporated into optimization mathematical models. Moreover, the simulation modeler's job does not stop when the development is completed, because validating the model's behavior and verifying its results are as important as developing the simulation model itself. Further discussion of these model verification and validation issues, however, is deferred until Chapter 4.

Computers are efficient simulation tools. During their emergence in the early 1950s, analyses of model designs were done via simulation by mapping the model designs onto the assembly language of the computers or onto a high level programming language like Fortran and measuring their performance. Not until

Geoffrey Gordon¹ introduced the General Purpose Simulation System (GPSS) in 1961 did computer simulation become a common tool, used by large corporations to answer “*What if...?*” questions without incurring millions of dollars in costs [56].

The three major types of computer simulations are continuous simulation, Monte Carlo simulation, and discrete event simulation [56]. In continuous simulations, the events are generated and evaluated continuously in time. In Monte Carlo simulations, events are generated randomly in time, according to some predefined probability distribution. In discrete event simulations, events are separated by blocks of time.

Since the market data collected is chiefly available on a yearly basis—i.e., discrete data—a discrete event simulation approach was chosen to model the dynamics of the supply of workstation components in this dissertation. The methodology is deterministic², and the blocks of time elapsed between each event are equal to one year. As the simulation evolves from year to year, the technological attributes of the components change in response to various technology-driving trends. The relationships between the component attributes and their technology-driving trends are presented in a set of relational diagrams. The relational diagram is the diagrammatic approach to displaying the model relations adopted here to help communicate the content of the component supply models.

¹Gordon developed GPSS at IBM.

²A deterministic system has no random values as input or output and, for each set of input data, there exists one output.

3.1.2 Relational Diagrams and Attributes

Figure 3.1 is an illustration of a relational diagram. In Figure 3.1, the state, or value, of an attribute at time t is represented as \mathbf{Y}_t and its initial state as \mathbf{Y}_0 . The index 0 indicates the year the simulation starts and the index t indicates a discrete year between the initial and the final years of the simulation period. Since the data is available on an annual basis, the simulation period of the study is given in number of years from the initial year, and the time index t is incremented by 1 every iteration of the model. The figure illustrates n technology-driving trends (TDTs), the relative changes of which are denoted by \mathbf{X}_{nt} s. Each technology-driving trend may acquire a new value every iteration, depending on the equation modeling its behavior. Only the relative change in the technology-driving trend's value at t (\mathbf{TDT}_t) with respect to its initial value at t_0 (\mathbf{TDT}_0) is needed to capture how it affects the attribute \mathbf{Y}_t . Hence, each \mathbf{X}_{nt} term is expressed as the ratio of the current and initial values of the technology-driving trend it represents:

$$\mathbf{X}_{nt} = \frac{\mathbf{TDT}_{nt}}{\mathbf{TDT}_{n0}}. \quad (3.1)$$

As illustrated in Figure 3.1, the attribute \mathbf{Y}_t is encircled in the center of the relational diagram, with its initial value \mathbf{Y}_0 in the upper half of the figure. The relative changes of the technology-driving trends' values, the \mathbf{X}_{nt} s, are shown in the lower half of the figure. \mathbf{Y}_0 and the \mathbf{X}_{nt} s are connected to the encircled \mathbf{Y}_t by directed edges, each with an edge weight label. The weights on each \mathbf{X}_{nt} edge, the \mathbf{a}_{ns} , are the exponents of these elements in the equation representing the behavior of the attribute. While tuning the values of the attribute's equation to match related actual market data, each \mathbf{a} is chosen to

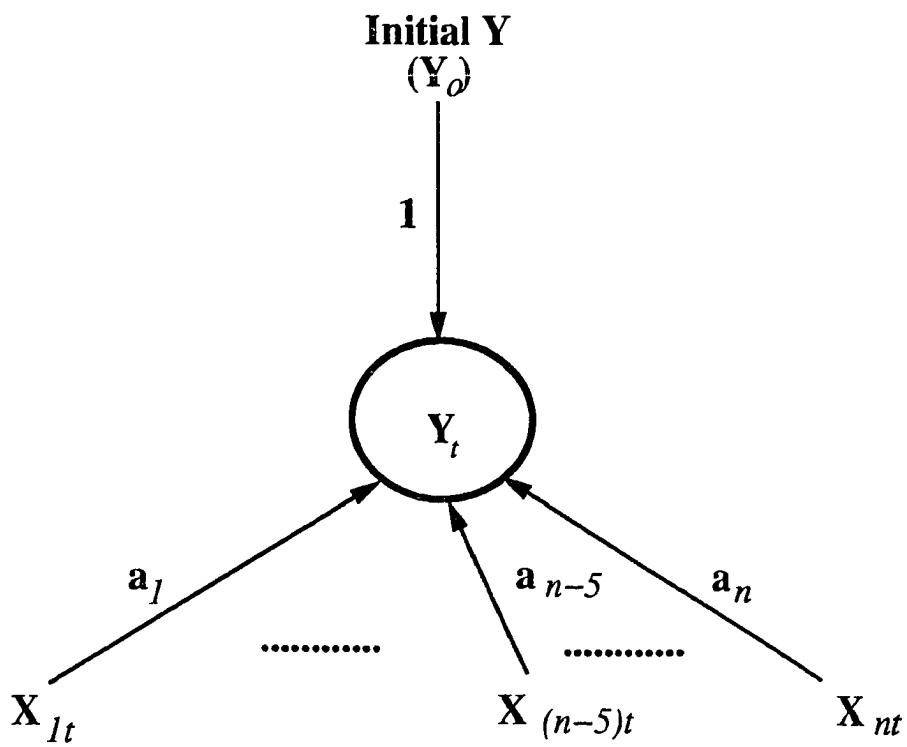


Figure 3.1: Illustration of a relational diagram.

reflect the degree to which each technology-driving trend affects, relatively, the attribute. If the absolute value of \mathbf{a}_n is larger than the absolute value of $\mathbf{a}_{n-\alpha}$, the effect of a change in \mathbf{TDT}_{nt} is considered more important to the overall value of the attribute than a change in $\mathbf{TDT}_{(n-\alpha)t}$. If \mathbf{a} is positive, an increase (or a decrease) in the technology-driving trend's value increases (or decreases) the value of the attribute; if \mathbf{a} is negative, an increase (or a decrease) in the trend's value decreases (or increases) the value of the attribute. The effect of the relative changes in the technology-driving trends' values on the attribute in the relational diagram can be expressed mathematically as:

$$\mathbf{Y}_t = \mathbf{Y}_0^{\mathbf{a}_0} * \mathbf{X}_{1t}^{\mathbf{a}_1} * \mathbf{X}_{2t}^{\mathbf{a}_2} * \dots * \mathbf{X}_{nt}^{\mathbf{a}_n} \quad (3.2)$$

$$= \mathbf{Y}_0^1 * \left(\frac{\mathbf{TDT}_{1t}}{\mathbf{TDT}_{10}}\right)^{\mathbf{a}_1} * \left(\frac{\mathbf{TDT}_{2t}}{\mathbf{TDT}_{20}}\right)^{\mathbf{a}_2} * \dots * \left(\frac{\mathbf{TDT}_{nt}}{\mathbf{TDT}_{n0}}\right)^{\mathbf{a}_n}. \quad (3.3)$$

If the partial derivative³ of \mathbf{Y}_t is computed with respect to one of the time-varying technology-driving trends, \mathbf{TDT}_{1t} for example, Equation 3.3 will be equal⁴ to:

$$\frac{\partial \mathbf{Y}_t}{\partial \mathbf{TDT}_{1t}} = \mathbf{Y}_0 * \mathbf{a}_1 * \frac{\mathbf{TDT}_{1t}^{\mathbf{a}_1 - 1}}{\mathbf{TDT}_{10}^{\mathbf{a}_1}} * \left(\frac{\mathbf{TDT}_{2t}}{\mathbf{TDT}_{20}}\right)^{\mathbf{a}_2} * \dots * \left(\frac{\mathbf{TDT}_{nt}}{\mathbf{TDT}_{n0}}\right)^{\mathbf{a}_n}. \quad (3.4)$$

If the positions of \mathbf{Y}_0 and \mathbf{a}_1 of the right-hand side of Equation 3.4 are switched, it will look as follows:

$$\frac{\partial \mathbf{Y}_t}{\partial \mathbf{TDT}_{1t}} = \mathbf{a}_1 * \mathbf{Y}_0 * \frac{\mathbf{TDT}_{1t}^{\mathbf{a}_1 - 1}}{\mathbf{TDT}_{10}^{\mathbf{a}_1}} * \left(\frac{\mathbf{TDT}_{2t}}{\mathbf{TDT}_{20}}\right)^{\mathbf{a}_2} * \dots * \left(\frac{\mathbf{TDT}_{nt}}{\mathbf{TDT}_{n0}}\right)^{\mathbf{a}_n}. \quad (3.5)$$

³The actual technology-driving trends data is chiefly available on a yearly basis—i.e., discrete data. However, continuous lines can be drawn through the discrete data, continuous functions can be used to model the behavior of the technology-driving trends over time, and derivatives of the continuous functions can be computed.

⁴The value of \mathbf{a}_0 in Equation 3.2 is equal to 1 because \mathbf{Y}_0 represents the value of the attribute at the first year of the study period, t_0 (if the values of all the \mathbf{X}_{nt} s affecting \mathbf{Y}_t are equal to 1, $\mathbf{Y}_t = \mathbf{Y}_0 = \mathbf{Y}_0^{\mathbf{a}_0}$; i.e., $\mathbf{a}_0 = 1$).

Moreover, when comparing Equation 3.5 with Equation 3.3, it is apparent the right-hand side of Equation 3.5 can be written as:

$$\frac{\partial \mathbf{Y}_t}{\partial \mathbf{TDT}_{1t}} = \mathbf{a}_1 * \frac{\mathbf{Y}_t}{\mathbf{TDT}_{1t}}. \quad (3.6)$$

By dividing both the left and the right sides of Equation 3.6 by the ratio $\frac{\mathbf{Y}_t}{\mathbf{TDT}_{1t}}$, Equation 3.6 becomes:

$$\frac{\frac{\partial \mathbf{Y}_t}{\partial \mathbf{TDT}_{1t}}}{\frac{\mathbf{Y}_t}{\mathbf{TDT}_{1t}}} = \frac{\frac{\partial \mathbf{Y}_t}{\mathbf{Y}_t}}{\frac{\partial \mathbf{TDT}_{1t}}{\mathbf{TDT}_{1t}}} = \mathbf{a}_1. \quad (3.7)$$

The middle fraction of Equation 3.7 can be approximated by the ratio of the fractional change of \mathbf{Y}_t and the fractional change of \mathbf{TDT}_{1t} :

$$\frac{\frac{\partial \mathbf{Y}_t}{\mathbf{Y}_t}}{\frac{\partial \mathbf{TDT}_{1t}}{\mathbf{TDT}_{1t}}} \simeq \frac{\frac{\Delta \mathbf{Y}_t}{\mathbf{Y}_t}}{\frac{\Delta \mathbf{TDT}_{1t}}{\mathbf{TDT}_{1t}}} \simeq \mathbf{a}_1. \quad (3.8)$$

From Equation 3.8, it is apparent that the fractional change in \mathbf{Y}_t —i.e. the ratio $\frac{\Delta \mathbf{Y}_t}{\mathbf{Y}_t}$ —can be approximated by the product of the technology-driving trend's fractional change and its exponent in Equation 3.3:

$$\frac{\Delta \mathbf{Y}_t}{\mathbf{Y}_t} \simeq \mathbf{a}_1 * \frac{\Delta \mathbf{TDT}_{1t}}{\mathbf{TDT}_{1t}}. \quad (3.9)$$

Thus, if the value of the technology-driving trend \mathbf{TDT}_{1t} were to change by 10%, the value of the attribute \mathbf{Y}_t changes by approximately 10% multiplied by the value of the technology-driving trend's exponent, \mathbf{a}_1 .

Relational Diagrams and Attributes: Example. It was pointed out in Chapter 2 that the CPU's feature size and die area affect its speed⁵ attribute.

⁵For example's sake, only two technology-driving trends are considered. For a complete development of the CPU speed attribute model, refer to Section 3.2.

Hence, if speed is denoted as Y_t in Equation 3.3 and feature size and die area are denoted as TDT_{1t} and TDT_{2t} , then the speed attribute equation might be expressed mathematically as:

$$SPEED_t = SPEED_0^{a_0} * \left(\frac{FeatureSize_t}{FeatureSize_0} \right)^{a_1} * \left(\frac{DieArea_t}{DieArea_0} \right)^{a_2}. \quad (3.10)$$

The value of a_0 in Equation 3.10 is equal to 1 because $SPEED_0$ represents the value of $SPEED_t$ in the first year of the study period, t_0 . The values of a_1 and a_2 are set while tuning the values of Equation 3.10 to match actual CPU speeds data over a certain period of study; the absolute value of a_2 ($|a_2|$) ought to be larger than the absolute value of a_1 ($|a_1|$) because die area has a relatively stronger influence on the CPU speed than does feature size. For example's sake, let us assume that the values of a_1 and a_2 are respectively set to -0.3 and 1 during the tuning process; the CPU speed attribute equation will look as follows:

$$SPEED_t = SPEED_0^1 * \left(\frac{FeatureSize_t}{FeatureSize_0} \right)^{-0.3} * \left(\frac{DieArea_t}{DieArea_0} \right)^1. \quad (3.11)$$

Equation 3.9 shows that a 5% decrease in the feature size is reflected by an approximate decrease of -1.5% ($-0.3 * 5\% = -1.5\%$), or increase of 1.5%, in the CPU speed. On the other hand, a 5% increase in the feature size creates an approximate increase of -1.5% ($-0.3 * 5\% = -1.5\%$), or decrease of 1.5%, in the CPU speed, assuming Equation 3.11 is correct.

Equation 3.9 shows also that a 5% increase in the die area is reflected in an approximate increase of 5% in the CPU speed, and similarly for a die area decrease. Note that the approximation in Equation 3.9 carries with it a canceling effect if an increase (or a decrease) occurs in both the feature size and the die area as a result of their exponents' opposite signs; for example, an increase (or

a decrease) of 5% in the feature size can be, approximately, canceled out by an increase (or a decrease) of 1.5% in the die area, and vice versa.

3.2 ICs Supply Model: Microprocessors and DRAMs

For the past four decades, capital intensity⁶ and short product life⁷ have been the chief characteristics of the semiconductor industry. By the year 2000, a new advanced semiconductor factory will probably cost about \$2 billion (10^9), and the United States will only be able to afford about 10 such factories [68]. In 1987 dollars, a fabrication laboratory cost approximately \$100 million (10^8), and required a 3% to 6% forecasted market share before it could be built [104]. Yet, despite the large capital outlays necessary to build a semiconductor factory, each semiconductor product may average no more than 6 months to 2 years of profitable sales before being obsoleted by new and enhanced products!

The main semiconductor components on the market today are microprocessors, microcontrollers, and memories. Many of these products have several generations of components already obsolete before them and most of them have one or two future versions under design and testing in the manufacturing laboratories. In this dissertation, the components examined are microprocessors and memories—in particular, CISC and RISC microprocessors and DRAMs. The CMOS semiconductor technology is given the greatest attention here.

This section is organized as follows: Subsection 3.2.1 presents historical

⁶In this context, capital intensity is synonymous with the large amount of capital required to set up an IC manufacturing plant.

⁷Short product life is synonymous with fast obsolescence rate.

data on the physical characteristics trends of ICs, Subsection 3.2.2 presents historical data on capabilities and price trends of ICs, Subsection 3.2.3 presents the assumptions and the terminology used in the development of the supply model of ICs, and Subsections 3.2.4 through 3.2.7 present the mathematical formulation of the supply model of CPUs and DRAMs.

3.2.1 Historical Data on the Physical Characteristics of ICs

If a company manufacturing ICs is to survive in this competitive and diversified market, its production must be characterized by high product yields. This criterion can only be satisfied with a clean manufacturing environment, constant attention to the latest technological developments, and learning⁸. Through learning, acquired techniques are transferred to new generation of products without the reinventing of the wheel.

Today, the Japanese fabrication laboratories are the most sophisticated and the cleanest in the world. The position of the Japanese ICs manufacturing firms on the learning curves and their clean environments have enabled them to achieve higher yields than their US and European counterparts. This allowed them to launch a DRAM dumping war in the 1980s and capture almost 90% of the world DRAM market [26].

As described in Chapter 2, the current technology for manufacturing ICs is very intricate, involving several masking levels and processing steps, thereby making the ICs quality and reliability very sensitive to a whole chain

⁸Learning is dependent on good documentation of the manufacturing processes.

of decisions during the production process. Adding to the production processes' intricacy are the ICs' functionalities and complexity. What follows is a presentation of ICs production trends⁹ including the increase of the die area, the reduction of the feature size, changes in the chip's configuration, the increase of the number of critical masks, and the increase of the silicon wafer diameter.

ICs Die Size Trends. Die size¹⁰ has steadily increased [48]. Figure 3.2 is a log-linear graph which illustrates the historical trend of increasing die areas for processors, ASICs, and memories, where the vertical axis indicates the base 10 logarithm of the chip's area¹¹ in square millimeters and the horizontal axis indicates the year of the ICs market introduction. Die areas of DRAMs and of Motorola's and Intel's CISC microprocessors can be followed in Figure 3.2. For instance, the die area of a 4 megabit DRAM, introduced in 1988, was close to 80 mm², and the die area of the Intel i80386 microprocessor, introduced in 1986, was close to 113 mm². Using the upper line shown in Figure 3.2 to relate the die area of a CISC microprocessor to its year of manufacture, the following equation results:

$$CISCDA_t = 10^{-0.1+0.0753*(t-1984)} \text{ (cm}^2\text{)} \quad (3.12)$$

where $CISCDA_t$ represents the CISC CPU die area in cm², and t represents the year the die area is calculated. Similarly, the lower line of Figure 3.2 can be used to relate the die area of a DRAM to its year of manufacture. This equation can

⁹Production trends refer to manufacturing trends and physical characteristics of the components. In this dissertation, they are called technology-driving trends.

¹⁰Die size and die area are synonymous in this context.

¹¹100 mm² = 1 cm².

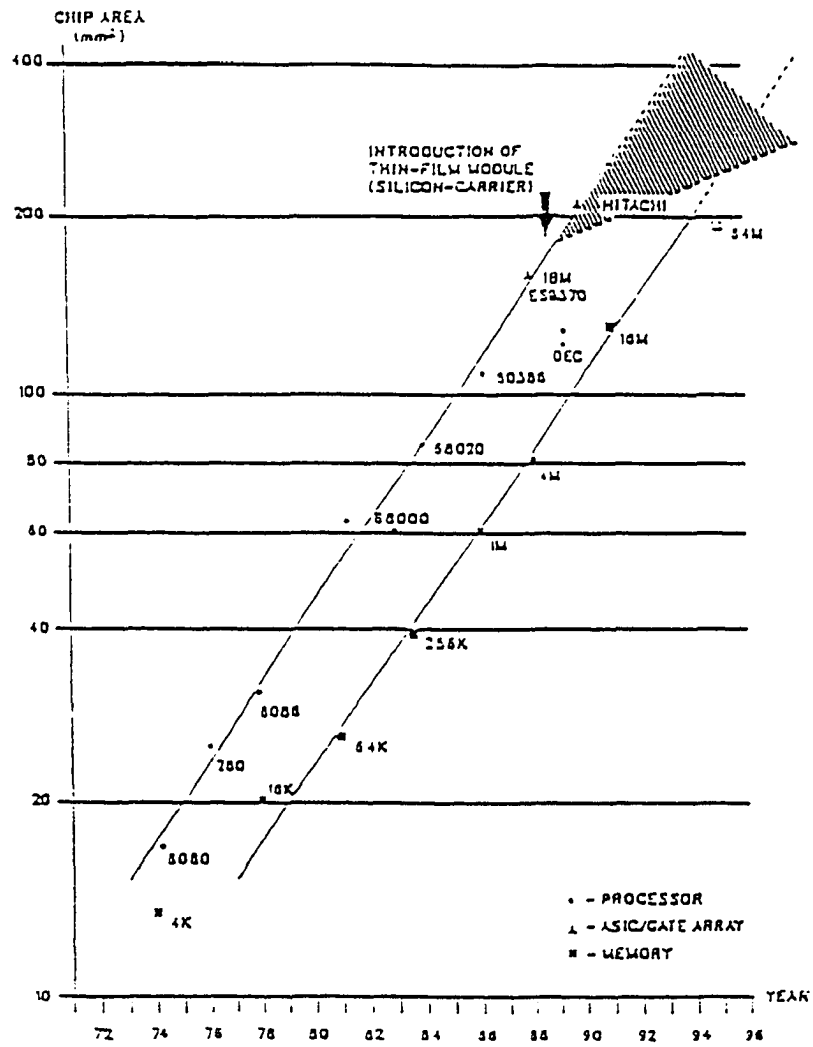


Figure 3.2: Die Sizes of CISC CPUs and DRAMs versus time. Source: [48].

be written as:

$$DRAMDA_t = 10^{-0.4+0.0753*(t-1984)} \text{ (cm}^2\text{)}. \quad (3.13)$$

The die area trends presented in Equations 3.12 and 3.13 will taper off in the future because of the manufacturing difficulties and the low yields associated with large area dies [33, 48, 71]. While the equations' values capture past trends, they do not project, with absolute certainty, future die areas. Nevertheless, they can be used to illustrate a "possible" future behavior of IC attributes. What follows is a presentation of the die yield as the main factor inhibiting the die area increase.

ICs Die Yields. Die yields data corresponding to different die areas was presented in Chapter 2 of Hennessy/Patterson [33]. The results, shown on Table 3.1, indicate that an increase in the die area induces a decrease in the die yield and, consequently, an increase in the cost of the die. A larger die has a larger area exposed to dust particles and manufacturing defects. Therefore, clean production environments and continuous production processes enhancements are essential for high yields. Since die yields are so crucial to the economics of producing ICs, the supply model will incorporate several factors affecting the die yield, including the die area, and generate the die yields for the corresponding die areas. (The model's die yield results will be presented in Subsection 4.3.1 and compared to the data provided in Table 3.1.)

ICs Feature Size Trend. The minimum ICs feature size has been decreasing since the early 1960s [92]. A smaller feature size is usually accompanied by a faster die operational speed, a higher die functionality, and an increase

ICs: Die Yields	
Die Area (cm ²)	Die Yield (Actual Data) (%)
0.0625	78
0.2601	46
0.5776	22
1.0404	10
1.6129	5
2.3104	3
3.1684	2
4.1209	1

Table 3.1: Die areas and their actual die yields. Source: [33].

in its manufacturing complexity and testing time. The historical trend of the minimum feature size is presented in Figure 3.3, a log-linear graph where the vertical axis indicates the base 10 logarithm of the feature size in microns and the horizontal axis indicates the time. On the graph, the feature size capabilities of different types of photoresists and lithography methods are included, each with time intervals representing the time of introduction and the time of obsolescence. For instance, the negative photoresist had been used with contact printing lithography from the early 1960s through the late 1970s. The line shown in Figure 3.3 can be expressed as the following equation:

$$FS_t = 10^{1.4 - 0.055 \cdot (t - 1960)} \text{ (micron)} \quad (3.14)$$

where FS_t represents the minimum feature size in microns, and t represents the year in which the feature size is calculated.

ICs Die Configuration Trend. Another factor that has an important effect on a die's functionality is its overall configuration. For example, new generation microprocessors are configured with an integer unit (IU), a floating point unit (FPU), and a memory management unit (MMU) on one single die, reducing the processing board's interconnection delays and improving the throughput. Other configurations dedicate a die to each of the IU, FPU, and MMU functions by using an enhanced packaging technology. In the 1970s, the packaging technology had to handle the high power consumption of bipolar ICs; however, the current semiconductor CMOS technology is characterized by very low power consumption. As a result, newer and more sophisticated packaging techniques

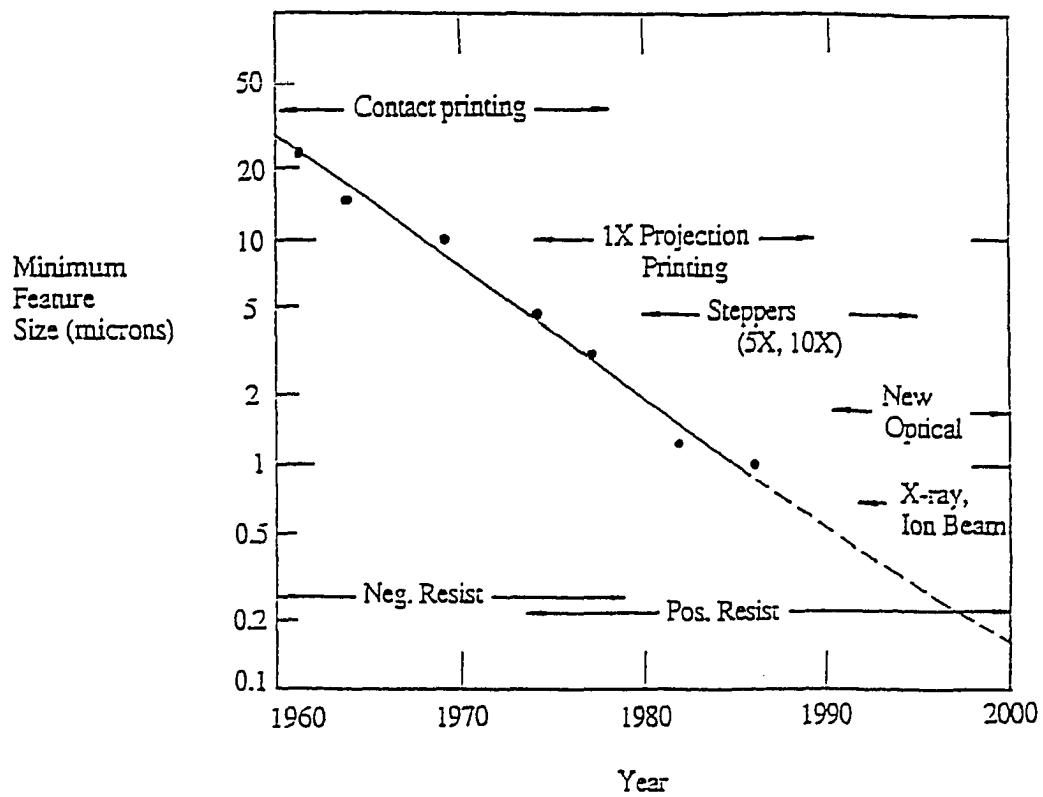


Figure 3.3: Feature size versus time. Source: [92].

have been developed, each with its physical limitations on the number of pins¹² the package can have and, correspondingly, the maximum number of watts¹³ it can consume [99]. Nevertheless, the maximum number of pins per CMOS package is far greater than for an equal area bipolar package when both chips are designed to operate at room temperatures.

Packaging is itself a highly developed field. Several new packaging technologies are discussed in [48, 66, 73, 95, 99]. One of the most promising new approaches is to distribute the dies on a silicon based thin film package, reducing communication and interconnection delays and allowing the dies' operational speeds to reach their respective limits without compromising performance with power consumption issues [48, 73].

ICs Production: Number of Critical Masks Trend. As presented in Section 2.2, during the production of an IC, several photoresist masks are used in the process of etching the circuit configuration and doping the corresponding etched areas with the specified impurities. The number of these masks depends on the functionality of the IC and the manufacturer's approach to achieving that functionality. Of these masks, a few are critical to the production of a good die and, consequently, its cost. The current average number of masks for most CMOS ICs is within the 14 to 20 range [18], and the number is projected to increase to 25 by the year 2000. The current average number of critical masks is within the 2 to 4 range [18, 33], or about 25% of the total number of masks.

¹²Package pins are leads connecting the die to its corresponding socket on the processing or memory board.

¹³Generated or consumed electric power is measured in watts.

If the number of critical masks stays constant as a fraction of the total number of masks, then, by the year 2000, it will grow to 6 to 8 in number. If we express this trend mathematically, the number of critical masks can be captured by the following equation:

$$CM_t = 25\% * [14.5 * 2^{0.08*(t-1990)}] \quad (3.15)$$

where CM_t represents the number of critical masks, and t represents the year in which the number of critical masks is calculated. Using Equation 3.15, the number of critical masks in 1980 is close to 2, a number that comports with actual data provided in [18, 33].

ICs Production: Silicon Wafer Diameter Trend. The trend in silicon wafer diameter, shown in Figure 3.4, has been toward increased size since the early 1960s [92]. The larger the wafer diameter is, the larger the number of dies per wafer and the cheaper the cost per die. The cost of the wafer increases as its area increases; however, the increase in the overall number of dies that can be produced from it offsets the wafer's cost increase. The log-linear graph of the historical trend of the wafer diameter, presented in Figure 3.4, has a vertical axis measuring the base 10 logarithm of the wafer diameter in millimeters and a horizontal axis indicating time. A line drawn through the points in Figure 3.4 to relate wafer diameter to time results in the following equation:

$$WD_t = 10^{1.48+0.0252*(t-1960)} \quad (mm) \quad (3.16)$$

where WD_t represents the silicon wafer diameter in millimeters, and t represents the year in which the wafer diameter is calculated.

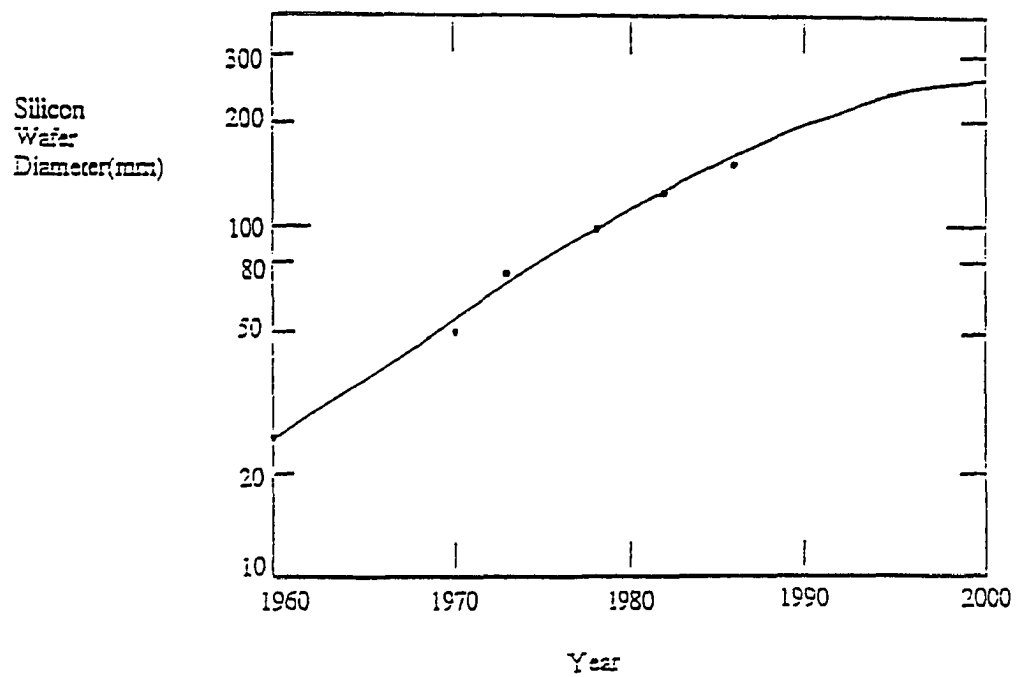


Figure 3.4: Silicon wafer diameter versus time. Source: [92].

Closure: When taken together, the historical trends equations—die area, feature size, number of critical masks, silicon wafer diameter—presented in the current subsection form a set that captures the ICs technology-driving trends as a function of time. These equations will be used as a part of the ICs supply model, to be presented later in this section.

3.2.2 Historical Data on ICs Capabilities and Price Trends

Actual data on ICs capabilities and prices were collected from several sources, including newspapers, magazines, journals, computer news network colleagues, and companies' data manuals. What follows are tabularized and graphed actual data of CISC and RISC CPUs and DRAMs.

CPUs Data. The following paragraphs present actual CISC and RISC CPUs price, performance, and physical dimension data, including CPU speed, MIPS, IPC¹⁴, and die area. The MIPS rating can be used as a *relative performance measure only to gain insight into the operational throughput of the CPU itself and not the whole computer system.*

CISC CPUs Data. Table 3.2 provides actual Intel and Motorola CISC CPUs data, including the CPU introduction date, its model number, its operational speed, its MIPS rating, its number of instructions per cycle (IPC) value, its die¹⁵ area in (mils x mils) and cm², and, finally, its speed per square centimeter rating.

¹⁴IPC is an acronym for the number of Instructions Per Cycle.

¹⁵1000 mils = 1 inch = 2.54 cm.

ICs: CISC CPUs							
Year	CPU Model #	Speed (MHz)	MIPS	IPC	Die Area (mils x mils)	Die Area (cm ²)	Speed/cm ²
1978	i8086	8	n.a.	n.a.	n.a.	0.28	28.6
1979	i8088	8	0.33	0.04	n.a.	n.a.	n.a.
1982	M68000	10	0.5	0.05	215x239	0.33	30.3
1982	i80286	16	2	0.125	n.a.	n.a.	n.a.
1983	M68010	10	0.6	0.06	239x267	0.41	24.4
1985	M68020	16.6	2	0.12	271x278	0.49	33.8
1985	i80386	16	3	0.187	n.a.	1.13	14.15
1987	M68020	20	3	0.15	271x278	0.49	40.8
1987	M68030	25	4	0.16	330x379	0.80	31.2
1987	i80386	20	4.3	0.215	n.a.	1.13	17.7
1988	i80386	25	6.1	0.244	n.a.	1.13	22.1
1988	i80386	33	8.3	0.25	n.a.	1.13	29.2
1989	i80486	25	11.4	0.45	414x619	1.65	15.2
1990	M68040	33	20	0.606	490x460	1.45	22.75
1990	i80486	33	15.2	0.46	414x619	1.65	20
1990	i80486	50	25	0.5	414x619	1.65	30.3
1991	i80486	66	35	0.53	414x619	1.65	40

Table 3.2: Actual market data of Intel and Motorola CISC CPUs. Sources: Intel data [41, 42, 48], Motorola data [61, 71, 84].

(Unavailable data is listed on the table as n.a.) Some of the listed processors perform integer operations only, and coprocessors perform the floating point and memory management operations.

The CPU's throughput is a function of its speed and its architecture. The speed reflects the sophistication of the manufacturing process and the feature size at which a CPU is manufactured. The architecture encompasses the instruction set used, CISC or RISC, and the number of instructions per cycle. The IPC values on Table 3.2 are obtained by dividing the MIPS rating by the speed of the corresponding CPU, and the speed per square centimeter values are obtained by dividing the speed of the corresponding CPU by its die area.

Table 3.3 provides actual Intel CISC CPUs prices and price per MIPS data over the 1979-1991 period. There is a difference, however, between the list price and the direct cost of producing an IC. The list price of an IC is usually equal to [33]:

$$List\ Price = \frac{Cost * (1 + Direct\ Cost\ \%)}{(1 - Gross\ Margin\ \%) * (1 - Average\ Discount\ \%)} (\$) \quad (3.17)$$

where the direct cost includes labor and product overhead costs; the gross margin incorporates the overall pretax profits, taxes, production costs (like R&D), marketing, sales, maintenance, loans interest payments, and rental payments [33]; and the average discount is used as an incentive for large volume buyers. The distinction between list price and direct cost is drawn because the supply models to be presented later will provide only the direct costs of the components, not the list price. To obtain the list price, a markup of up to 200% to 300% is necessary, depending on the ICs manufacturer and the product [33]. Each company might have different percentages for the direct costs, the gross

ICs: CISC CPUs			
Year	Model #- Speed (MHz)	Price (Actual Data) (\$)	Price/MIPS (Actual Data) (\$)
1979	i8088-8	360	1091
1982	i80286-16	360	180
1985	i80386-16	299	99.6
1989	i80486DX-25	700	61.4
1991	i8086-8	1.5	n.a.
1991	i80286-16	7	3.5
1991	i80386DX-33	166	20
1991	i80486DX-66	588	16.8
1991	i80486SX-25	258	22.6

Table 3.3: Actual prices and price/MIPS market data of Intel CISC CPUs.
Sources: [41, 42].

margin, and the average discount, and the values of their percentages depend on the product's attributes and quality, the market competition, and the overall market demand for similar products.

Actual RISC CPUs Data. Table 3.4 presents speeds, MIPS ratings, and IPCs of the Hewlett-Packard Precision Architecture (HP-PA) and the Sun Microsystems Scalable Processor Architecture (SPARC) RISC machines. No CPU price data was available, and most of the die sizes are still considered as proprietary company information. As far as the IPC values are concerned, it is apparent that the RISC IPCs are larger than the CISC (for more details refer back to Section 2.2).

Actual Data for CISC and RISC CPUs Instructions per Cycle. Log-linear graphs of Intel and Motorola CISC CPU IPCs, and of HP and Sun RISC CPU IPCs are presented in Figures 3.5 and 3.6, respectively, where each vertical axis indicates the base 10 logarithm of the IPC of the corresponding processor and each horizontal axis indicates the time. The CISC IPC trend in Figure 3.5 can be expressed by the following equation:

$$CIPC_t = 0.05 * 10^{0.11*(t-1982)} \quad (3.18)$$

and the RISC IPC trend in Figure 3.6 can be expressed by the following equation:

$$RIPC_t = 0.6 * 10^{0.063*(t-1987)} \quad (3.19)$$

If extrapolated, the trend lines suggest an ever increasing IPC is possible over time. In fact, the IPC may be limited to upper bound values in the 3 to 4 range,

ICs: RISC Machines							
Year	System	Speed (MHz)	MIPS	IPC	Die Area (milsxmils)	Die Area (cm ²)	Speed/cm ²
1987	HP9000/825	12.5	9	0.72	n.a.	n.a.	n.a.
1987	Sun 4/260	16.67	10	0.599	n.a.	n.a.	n.a.
1988	HP9000/835	15	14	0.93	n.a.	n.a.	n.a.
1988	Sun 4/110	14.28	7	0.49	n.a.	n.a.	n.a.
1989	HP9000/845	30	22	0.73	551x551	1.96	15.3
1989	Sparcstation1	20	12.5	0.62	n.a.	n.a.	n.a.
1990	Sparcstation1+	25	15.8	0.63	n.a.	n.a.	n.a.
1990	Sparcstation2	40	28.5	0.71	n.a.	n.a.	n.a.
1991	HP9000/720	50	59	1.14	551x551	1.96	25.5
1991	HP9000/730	66	76	1.15	551x551	1.96	33.67

Table 3.4: Actual performance data of HP-PA and Sun SPARC RISC machines.
Sources: HP data [7, 22, 37, 51, 53, 89, 103], Sun data [84, 86].

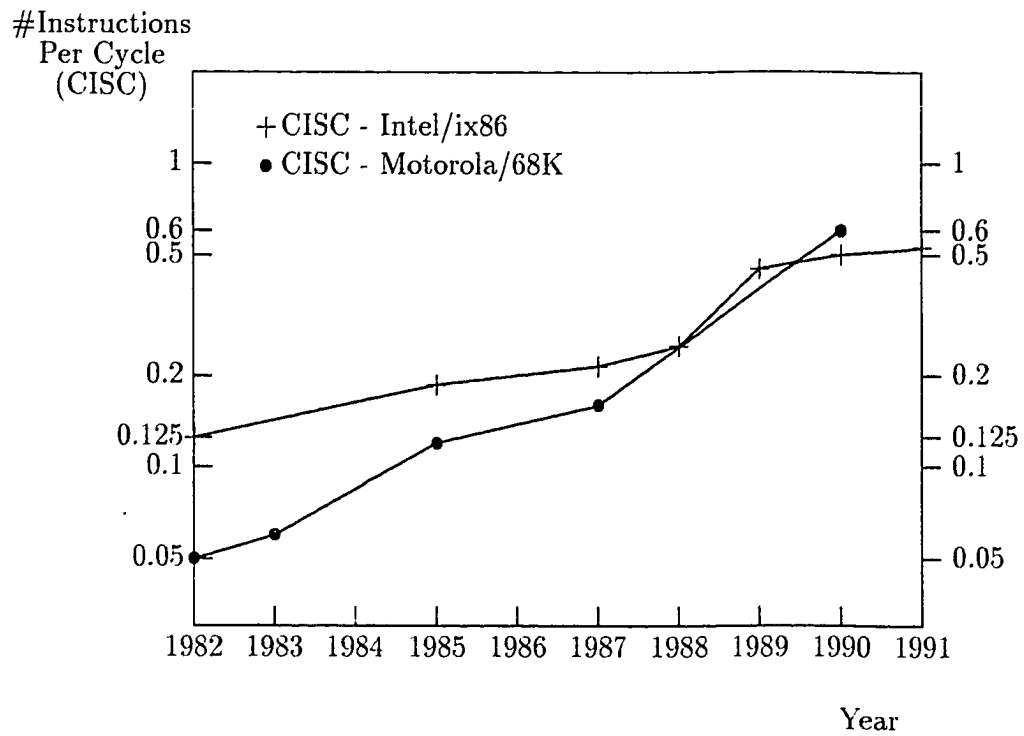


Figure 3.5: Actual data on the number of instructions per cycle for Intel and Motorola CISC CPUs. Sources: Intel data [41, 42, 48], Motorola data [61, 71, 84].

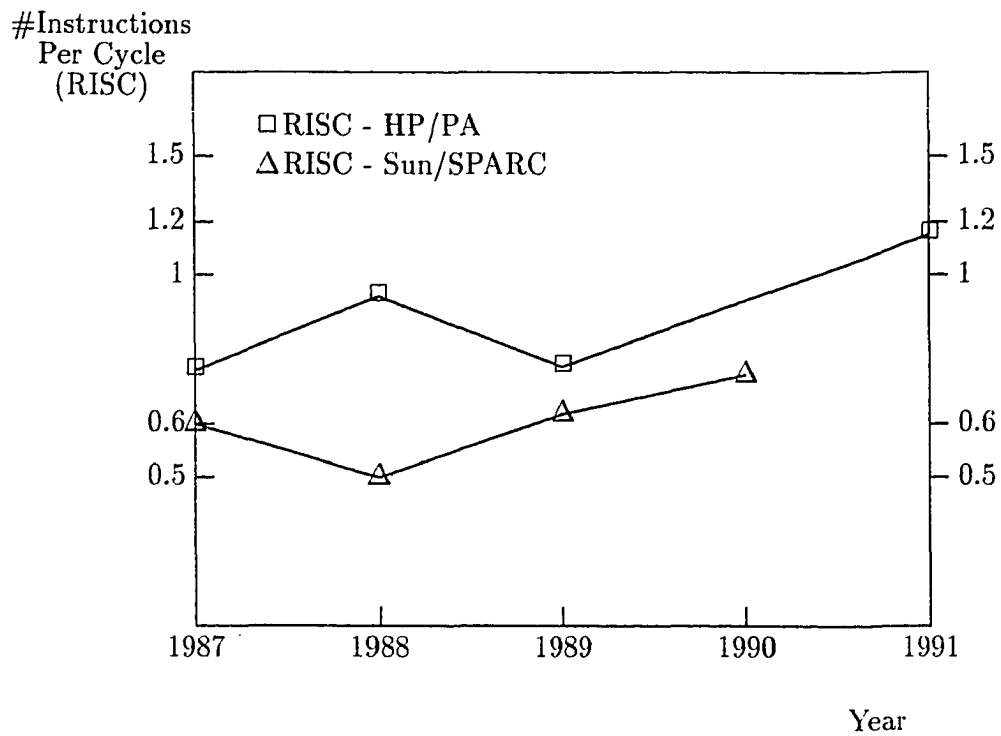


Figure 3.6: Actual data on the number of instructions per cycle for HP-PA and Sun SPARC RISC CPUs. Sources: HP data [7, 22, 37, 51, 53, 89, 103], Sun data [84, 86].

which might be reached by—if not before—the year 2000 [71, 97]. The limit is set by the architectural designs and the frequency at which a branch instruction is executed. The architectural designs limitations were discussed in Section 2.2. The branching limitations are related to the maximum number of instructions that can be executed in one clock cycle.

Each processor's assembly language has a set of branch instructions which are used to move around the program counter¹⁶ in the object code during execution. The program counter's branch destination depends on the type of the branch and the logical condition, if any, that the branch instruction must satisfy before being acknowledged. It has been estimated that there are 3 to 4 instructions between two branch instructions, so that up to four instructions might be executed in one cycle ($IPC = 4$) before the branch occurs and another set of instructions can be executed [71]. This practical limitation in IPCs needs to be kept in mind, since it will influence future projected trends of CPU MIPS or any other CPU performance rating.

DRAMs Data. Table 3.5 provides actual DRAM data. Shown for each DRAM are its capacity in bits, its introduction date, its die size in cm^2 , its density in number of megabits per square centimeter (Mb/cm^2), and its number of megabytes per square centimeter (MB/cm^2). The Mb/cm^2 data was obtained by dividing the K-indexed capacity values by the product of the corresponding die area and the number 1000, and the M-indexed capacity values by the corre-

¹⁶A program counter is the processor's hardware pointer to the program being executed.

ICs: DRAMs				
Year	DRAM (bits)	Die Area (cm ²)	Mb/cm ²	MB/cm ²
1978	16K	0.2	0.08	0.01
1981	64K	0.24	0.27	0.034
1984	256K	0.4	0.64	0.08
1986	1M	0.6	1.67	0.21
1988	4M	0.8	5	0.625
1991	16M	1.34	11.94	1.49

Table 3.5: Actual market data on DRAM die sizes, capacities, and densities.
Source: [48].

sponding die area. The MB/cm² was obtained by dividing the Mb/cm² values¹⁷ by 8.

Actual Motorola DRAM prices and price per megabyte data is presented on Table 3.6. Since the mid-1980s, DRAM prices decreased much faster than the manufacturers expected due to fierce Japanese competition [26]. Again, it is emphasized that the prices shown are list prices, not direct costs, and are marked up over costs by as much as 300% [33].

3.2.3 Model Assumptions and Terminology

The ICs supply model incorporates equations capturing the attributes of CMOS CPUs and DRAMs. The attributes considered are the costs, speeds, and MIPS of CPUs, and costs and capacities of DRAMs. Before presenting the mathematical formulation of the model, several assumptions are listed. This is followed by a list defining the variables and the parameters used in the model.

Model Assumptions. A relatively short period of study—1980 through 1996—is adopted because of the rapid pace at which the computer industry is changing and the uncertainties associated with it. Choosing a short period makes possible other assumptions listed below:

- Technological trends of the past in overcoming physical manufacturing barriers continue during the period of study.

¹⁷8 bits are required to form 1 byte.

ICs: DRAMs			
Year	DRAM (bits)	Price (Actual Data) (\$)	Price/MB (Actual Data) (\$)
1984	16K	1.09	545
1984	64K	3.4	425
1984	256K	17.9	559.3
1986	1M	100	800
1988	4M	264	528
1991	16M	329	164.5

Table 3.6: Actual prices and price/megabyte market data of Motorola DRAMs.
Source: [62].

- Past trends in IC manufacturing yields will continue to increase or, at worst, remain constant.
- Past trends in IC testing will continue to improve to deal with higher density and higher complexity chips, thereby improving the reliability of the ICs.

Definitions and Terminology. In presenting the model's equations, several parameters and variables are introduced, and they are defined as follows:

0	the time index for the first year of the study period
t	the time index
FS	the minimum feature size, in microns
CM	the number of critical masks
WD	the wafer diameter, in centimeters
WC	the wafer cost, in dollars
DA	the die area, in cm^2
DPW	the number of dies per wafer
$TDPW$	the number of test dies per wafer
DY	the die yield, in %
WY	the wafer yield, in %
FTY	the final die test yield, in %
$DPUA$	the number of silicon wafer defects per unit area
$ADTT$	the average die test time, in seconds
$TCPH$	the die testing cost per hour, in dollars
TC	the die testing cost, in dollars

<i>MC</i>	the die manufacturing cost, in dollars
<i>PC</i>	the die packaging cost, in dollars
<i>BIC</i>	the IC burn-in cost, in dollars
<i>ICTC</i>	the IC total cost, in dollars
<i>SPEED</i>	the speed of operation of the IC, in megaHertz
<i>SPEEDPER</i>	the speed of operation of the IC per square centimeter
<i>CIPC</i>	the number of CISC instructions per CPU clock cycle
<i>RIPC</i>	the number of RISC instructions per CPU clock cycle
<i>CMIPS</i>	the number of million CISC instructions per second
<i>RMIPS</i>	the number of million RISC instructions per second
<i>MEMORY</i>	the capacity of a DRAM, in megabytes
<i>MEMPER</i>	the number of DRAM megabytes per square centimeter

The supply model of ICs is formulated as follows: Subsection 3.2.4 presents the CISC and RISC CPU speed and MIPS models, Subsection 3.2.5 presents the DRAM capacity model, Subsection 3.2.6 presents the time behavior of dies of fixed capability (fixed CPU speed and fixed DRAM capacity ICs), and Subsection 3.2.7 presents the IC cost model with a description of the die manufacturing cost, the testing cost, the packaging cost, the burn-in cost, and the associated die yields all through the IC manufacturing process.

3.2.4 CPU Speed and MIPS Models

The MIPS rating is used in the model as a relative measure that reflects on the operational speed behavior of the CPU and its architectural enhancements over the years, where the architectural enhancements can be captured by the IPC

trends provided earlier. Since there are two different instruction set computers, CISC and RISC, two MIPS equations are formulated accordingly:

$$CMIPS_t = SPEED_t * CIPC_t \quad (3.20)$$

$$RMIPS_t = SPEED_t * RIPC_t. \quad (3.21)$$

The relational diagram¹⁸ for the operational speed model is illustrated in Figure 3.7. It relates the CPU speed at t to the initial CPU speed at t_0 and the technology-driving trends of feature size, die area, and number of critical masks:

- Feature size was chosen as a CPU speed technology-driving trend because a smaller feature size reduces the space between the chip's wires and modules, hence reducing the distance the electrons have to travel and increasing the die speed.
- Die area was chosen as a CPU speed technology-driving trend because a larger die area makes possible an increase in the functionality of the CPU. Several specialty ASICs that had been implemented outside the CPU, like the memory management unit and the floating point unit, can be included in the larger CPU area, thus eliminating the speed limiting communication outside the CPU and enabling a CPU's speed to reach limits unattainable through the packaging and early manufacturing technologies.
- Number of critical masks was chosen as a CPU speed technology-driving trend because it reflects the degree of manufacturing sophistication and

¹⁸Refer to Section 3.1 for a description of the relational diagram symbols and their interpretations.

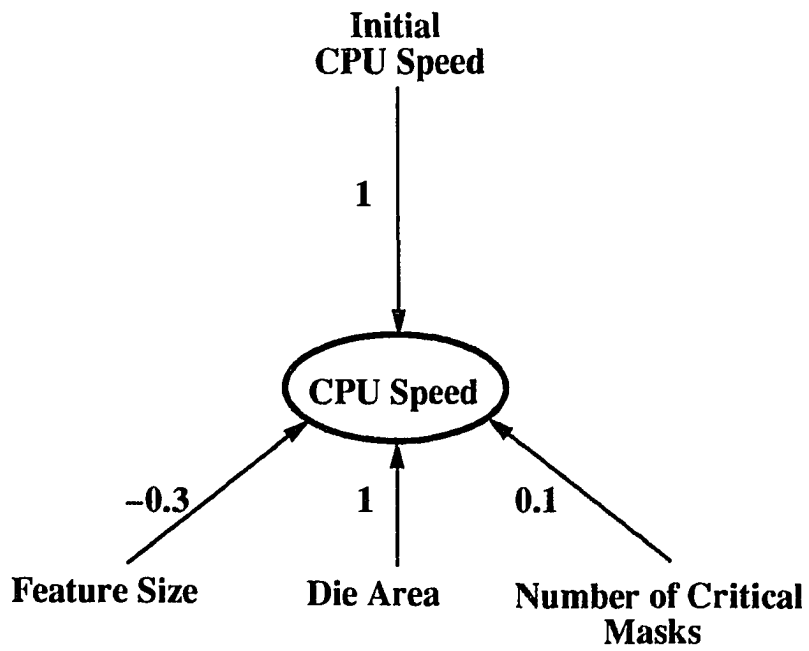


Figure 3.7: Relational diagram of the operational speed of CPUs.

the level of integration of the CPU production process. Both factors affect the CPU speed in the long run.

The CPU speed model can be expressed mathematically as follows:

$$SPEED_t = SPEED_0 * \left(\frac{FS_t}{FS_0}\right)^{-0.3} * \frac{DA_t}{DA_0} * \left(\frac{CM_t}{CM_0}\right)^{0.1} \quad (MHz). \quad (3.22)$$

It indicates that the CPU speed, at time t , increases as the feature size decreases and as the die area and the number of critical masks increase. The values of the exponents (a_n s) in Equation 3.22 were obtained by tuning the model to yield results to match the actual CPU speeds data provided in Subsection 3.2.2. All of the $|a_n|$ s are ≤ 1 , and each reflects an estimate of its relative importance to the behavior of the CPU speed attribute over time. For instance, the exponent of the die area element— $a_2 = 1$ —in Equation 3.22 is larger than the number of critical masks’— $a_3 = 0.1$ —because the die area plays a more important role in determining the CPU speed than the number of critical masks.

If the sign of an exponent (a_n) is positive, an increase (or a decrease) in the corresponding technology-driving trend’s value increases (or decreases) the CPU speed; if the sign is negative, an increase (or a decrease) in the trend’s value decreases (or increases) the CPU speed. For example, a 10% decrease in the feature size leads to an approximate increase of 3% in the CPU speed (see Equation 3.9), while a 10% increase in the die area leads to an approximate increase of 10% in the CPU speed, and a 10% increase in the number of critical masks leads to an approximate increase of 1% in the CPU speed.

3.2.5 DRAM Capacity Model

The speed of DRAM ICs is critical to a computer system's performance. However, since they are needed in large quantities for the system to have high performance measures, their interconnection and communication delays can offset their operational speed. Fast memory chips with higher bit densities are more desirable than an equal capacity of interconnected chips with a lesser bit density.

The DRAM IC is not as complex as a CPU IC because several processing functions are not implemented onto a single die. Rather, the DRAM has a simple structure formed by a repetition of memory cells, each separated by a pitch usually equal to the feature size. The number of critical masks is essentially equal to that of the CPUs' [18].

The relational diagram of the DRAM capacity model is illustrated in Figure 3.8. It relates the DRAM capacity at t to the initial DRAM capacity at t_0 and the technology-driving trends of feature size, die area, and number of critical masks:

- Feature size was chosen as a DRAM capacity technology-driving trend because a smaller feature size reduces the space between the memory cells and the space each of them occupies on the die, thus increasing the DRAM capacity.
- Die area was chosen as a DRAM capacity technology-driving trend because a larger die can incorporate more memory cells and have a larger DRAM capacity.

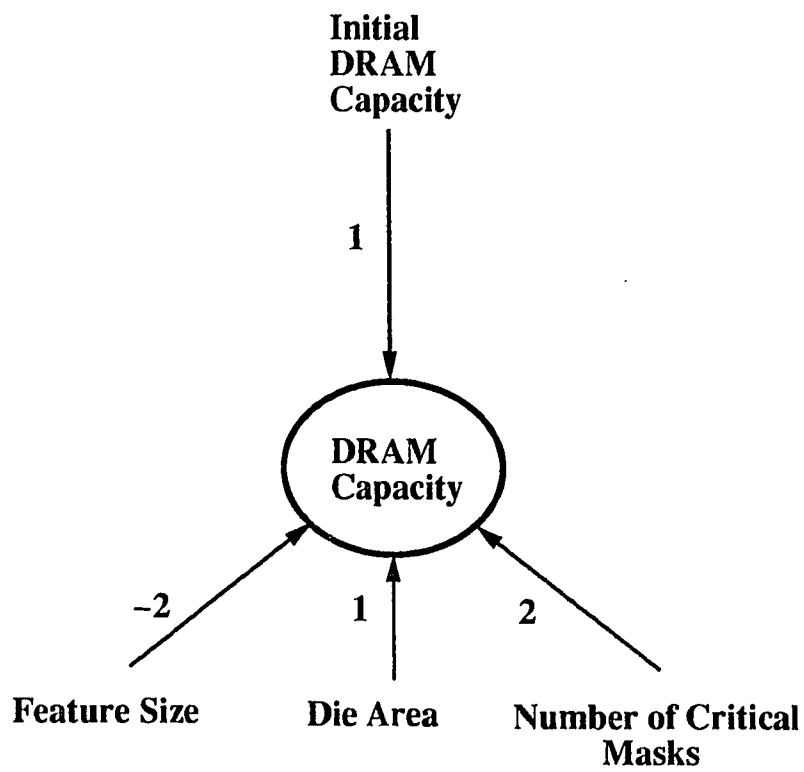


Figure 3.8: Relational diagram of the DRAM capacity.

- Number of critical masks was chosen as a DRAM capacity technology-driving trend because it reflects more sophisticated DRAM manufacturing processes where more than one memory cell can occupy the same space only one cell occupied in an earlier DRAM layout.

The DRAM capacity model can be expressed mathematically as follows:

$$MEMORY_t = MEMORY_0 * \left(\frac{FS_t}{FS_0}\right)^{-2} * \frac{DA_t}{DA_0} * \left(\frac{CM_t}{CM_0}\right)^2 \quad (MB). \quad (3.23)$$

The DRAM capacity increases as the feature size decreases and as the die area and the number of critical masks increase. The values of the exponents (a_n s) in Equation 3.23 were obtained by tuning the model to yield results to match the actual DRAM capacity data provided in Subsection 3.2.2. All of the $|a_n|$ s are ≥ 1 , and each reflects an estimate of its relative importance to the behavior of the DRAM capacity attribute over time. For instance, a 10% decrease in the feature size leads to an approximate increase of 20% in the DRAM capacity (see Equation 3.9), and similarly for the other factors of Equation 3.23.

3.2.6 Die Areas for Fixed Capability ICs

For a fixed functionality or attribute IC (i.e., a fixed CPU operational speed in the case of microprocessors, or a fixed DRAM capacity in the case of memories), the die size decreases over time—this is because the minimum feature size decreases and the value of the number of critical masks increases. Decreasing die areas allow increasing die yields and decreasing die costs. The corresponding die area equations for each type of fixed capability IC can be derived from Equations 3.22 and 3.23. They are:

- Fixed Speed CPU

$$DA_t = \frac{\text{Fixed SPEED}}{\text{SPEEDPER}_0} * \left(\frac{FS_t}{FS_0}\right)^{0.3} * \left(\frac{CM_t}{CM_0}\right)^{-0.1} \quad (cm^2) \quad (3.24)$$

where SPEEDPER_0 is equal to the initial CPU speed at t_0 (SPEED_0) divided by its corresponding die area at t_0 (DA_0).

- Fixed Capacity DRAM

$$DA_t = \frac{\text{Fixed DRAM}}{\text{MEMPER}_0} * \left(\frac{FS_t}{FS_0}\right)^2 * \left(\frac{CM_t}{CM_0}\right)^{-2} \quad (cm^2) \quad (3.25)$$

where MEMPER_0 is equal to the initial DRAM capacity at t_0 (MEMORY_0) divided by its corresponding die area at t_0 (DA_0).

3.2.7 IC Cost Model

The total direct manufacturing cost of a packaged and operational IC is equal to:

$$ICTC_t = \frac{MC_t + TC_t + PC_t + BIC_t}{FTY_t} \quad (\$) \quad (3.26)$$

where MC_t represents the IC manufacturing cost, TC_t the testing cost, PC_t the packaging cost, BIC_t the burn-in cost, and FTY_t the final test yield. No IC indirect costs are considered in Equation 3.26, as they are assumed as internal company decisions. What follows are descriptions of each of the cost factors of Equation 3.26, each with its own mathematical expression.

IC Manufacturing Cost. The die manufacturing cost depends on the silicon wafer cost, the manufacturing die yield, and the number of manufacturable dies

per wafer. The manufacturing cost equation is:

$$MC_t = \frac{WC_t}{DPW_t * DY_t} \quad (\$) \quad (3.27)$$

where WC_t is the silicon wafer cost, DPW_t is the number of dies per wafer, and DY_t is the die yield. The paragraphs below discuss these components of Equation 3.27 in turn.

Silicon Wafer Cost. Silicon ingot manufacture has improved over the years and, consequently, the quality of the wafers sliced from the ingots has also increased. The yield of good and processable silicon wafers per ingot has reached the 80% to 90% level during the past decade [18, 33], while, at the same time, the overall cost of a fixed diameter wafer has dropped. However, as the wafer diameters have increased, so have their costs, offsetting some of the cost reductions made possible by improved manufacturing.

The only data available on the wafer cost is in [33], where a wafer, in 1990, cost from \$500 to \$550. From Equation 3.16, the wafer diameter in 1990 was 16.25 centimeters. Unfortunately, no wafer price data for other years were available, so no price trend could be estimated from historical data. Based on conversations with colleagues, we simply assumed that wafer costs have increased by \$20 per year since 1980, which is represented by the equation:

$$WC_t = 350 + (t - 1980) * 20 \quad (\$). \quad (3.28)$$

Actual wafer cost data varies from one manufacturer to another, and most companies consider it proprietary information. It is assumed that the \$20 per year increase reflects the net effects of diameter increase, inflation, and the improved manufacturing yields.

Number of Dies per Wafer. Since dies are rectangular and wafers are usually round, the partitioning of the wafer results in some round edges that cannot be manufactured into the specified die dimensions. Furthermore, most companies select two or more dies per wafer for testing purposes, which further reduces the area of the wafer used for packable ICs. The number of dies per wafer is expressed as:

$$DPW_t = \frac{\pi * WD_t^2}{4 * DA_t} - \frac{\pi * WD_t}{\sqrt{2} * DA_t} - TDPW_t \quad (3.29)$$

where the first term is equal to the wafer area divided by the die area, the second term is the wasted round wafer area, and the third term is the number of testing dies per wafer. Usually the number of testing dies per wafer is equal to 2 [33].

Die Yield. By increasing the die area, three consequences ensue: the number of dies per wafer becomes smaller, the cost per die increases (Equation 3.27), and the die yield decreases [33]. Several manufacturing die yield approximations have been formulated [19, 33], and the one chosen here was first presented in Chapter 2 of Hennessy/Patterson [33]:

$$DY_t = WY_t * \left(1 + \frac{DPU A_t * DA_t}{CM_t}\right)^{-CM_t} \quad (3.30)$$

Remember that the die yield factor appears in the denominator of Equation 3.27, the expression for the manufacturing cost. When averaging the costs of the faulty and good dies, a smaller die yield will lead to a higher cost for each good die. Generally, the die yield has been a major factor in limiting the increase of the average manufacturable die area [33].

Due to the number of masks needed to produce a final chip, the die area could be exposed, at each mask, to etching and dust particles errors and

faulty connections. These errors are translated into die defects and cause the elimination of the die from the packaging stage. Equation 3.30 reflects the relative importance of the number of critical masks (CM), the die area (DA), and the number of silicon wafer defects per unit area (DPUA) on the manufacture of dies.

The wafer yield (WY) is included in Equation 3.30 because a faulty die could result from a wafer manufacturing defect which was not detected until the die production process. Wafer yields range from 80% to 90% [18, 33].

Number of Silicon Wafer Defects per Unit Area. The number of silicon wafer defects per unit area is influenced by the feature size's decreasing trend, the increasing trends over time of the IC's number of critical masks and die area, and the manufacture of the silicon wafer. Since each company guards its historical defects per unit area trends as classified information, no actual data was available. Because the improving position of the IC and wafer manufacturers on the learning curves has tended to offset the decrease of the feature size and the increase of the die area and the number of critical IC production masks, the number of silicon wafer defects per unit area was assumed to stay constant at 2.5 defects/cm² during the period of study. This assumption will be modified as necessary when additional data becomes available.

Manufactured Die Testing Cost. Each manufactured die is put through a series of tests to check its functionality and operational behavior. In some cases, mostly in CPU ICs, more than one function is implemented on a die. So, if one part of the die performing one of its functions is shown by the test to be

faulty, this part can be disabled and the CPU sold with the disabled function deleted as one of its attributes. In fact, Intel has done just this with its latest series of i80486 microprocessors. One of the i80486 CPU's functions handles the execution of floating point operations; if the floating point unit (FPU) is operational, the CPU is labeled i80486DX and sold for \$588 (Table 3.3). If it is faulty, the CPU with a disabled FPU is labeled i80486SX and sold for \$258. The moral of the story is that, with increasing die areas and die functionalities, it is costly to throw away a die if only a portion of it is faulty and can be disabled.

With each die test performed there is an associated cost. This cost is a function of the testing cost per hour (TCPH) at the manufacturing facility, the average die testing time (ADTT), and the yield of the dies that pass the test (DY). The testing cost equation is:

$$TC_t = \frac{TCPH_t * ADTT_t}{DY_t * 3600} (\$) \quad (3.31)$$

where the testing cost per hour depends on the manufacturer and the testing equipment used. The following paragraphs explain each of these factors.

Die Testing Cost per Hour. Currently, the die testing cost per hour ranges from \$200 to \$300 [18]. We assume its value in 1990 was \$240/hour and increasing at the rate of \$3 per year. The rate of increase is arbitrary because it depends on where the manufacturer falls on the learning curves—data that is unavailable to the public due to the competitive nature of the business. These assumptions, when expressed mathematically, yield:

$$TCPH_t = TCPH_{1990} + 3 * (year - 1990) (\$/hour). \quad (3.32)$$

Average Die Testing Time. The relational diagram of the average die testing time is illustrated in Figure 3.9, where the average die testing time at t ($ADTT_t$) depends upon the initial average die testing time at t_0 ($ADTT_0$), the feature size, the die area, and the number of critical masks:

- Feature size was chosen as an average die testing time technology-driving trend because a smaller feature size increases the number of transistors on and the functionality of a chip (as seen in Section 2.2), thus increasing the number of modules to test and, in the process, increasing the die testing time.
- Die area was chosen as an average die testing time technology-driving trend because a larger die can incorporate more modules and more application specific functions, thus increasing the amount of time to test it.
- Number of critical masks was chosen as an average die testing time technology-driving trend because it reflects the degree of manufacturing sophistication and the level of integration the CPU production process has reached; both factors increase the complexity of the chip and the time to test it.

The average die testing time can be expressed mathematically as follows:

$$ADTT_t = ADTT_0 * \left(\frac{FS_t}{FS_0}\right)^{-0.3} * \frac{DA_t}{DA_0} * \left(\frac{CM_t}{CM_0}\right)^{0.3} \quad (sec). \quad (3.33)$$

The average die testing time increases as the feature size decreases and as the die area and the number of critical masks increase. The values of the exponents (a_n s) in Equation 3.33 were obtained by tuning the model to yield results to fall

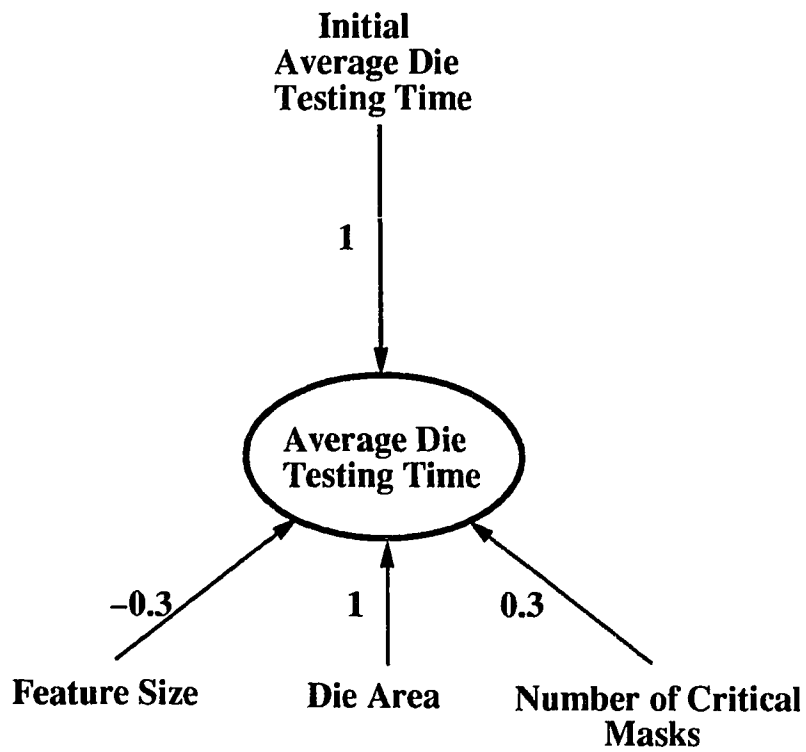


Figure 3.9: Relational diagram of the average IC die testing time.

into the testing time ranges observed historically. As of 1991, the average die testing time took anywhere from 10 seconds to 1 minute [18], depending on the die's complexity. Though hard data are not available, the average die testing time in 1980 was assumed to be in the 15 seconds to the 1.5 minutes range.

All of the $|a_n|$ s in Equation 3.33 are ≤ 1 , and each reflects an estimate of its relative importance to the behavior of the average die testing time over time. For example, a 10% decrease in the feature size leads to an approximate increase of 3% in the average die testing time (see Equation 3.9), and similarly for the other factors of Equation 3.33.

IC Packaging Cost. The IC packaging cost depends on the die area, the number of pins per package, and the type of package [33]. The type of package used is dependent on the amount of power the die consumes and on the frequencies at which it is operated. For die areas less than 1.1 cm^2 , the number of pins¹⁹ usually ranges from 100 to 200, and a plastic package is recommended because the heat dissipated is within the limitations of the plastic used [33]. Even though several new plastic packages with high heat tolerance ratings are available, ceramic packages are recommended [33] for die areas in the $1.1+$ cm^2 range, because they can better handle the heat generated from the power dissipated in the increased die area and the higher number of package pins. Packages with up to 1000 pins can be manufactured [99], but for the ICs' within the scope of this dissertation, a package with up to 500 pins is more than adequate. For a die area less than

¹⁹The number of pins increases with the functionality of the die.

1.1 cm², the packaging cost is [33]:

$$PC_t = 5 \text{ (\$)} \text{ for } DA \leq 1.1\text{cm}^2 \quad (3.34)$$

and for a die area larger than 1.1 cm², the packaging cost is [33]:

$$PC_t = 52 \text{ (\$)} \text{ for } DA \geq 1.1\text{cm}^2. \quad (3.35)$$

We assume that the packaging costs remain constant over the study period of the model simulation.

IC Burn-in Cost. After packaging, the die goes through a burn-in phase, which costs about 25 cents [33]:

$$BIC_t = 0.25 \text{ (\$)}. \quad (3.36)$$

The burn-in cost is assumed to be constant over the model's period of study.

IC Final Test Yield. The final test yield of good dies, after burn-in, is incorporated as the denominator in the IC total cost equation (Equation 3.26), making the good dies pay for the faulty ones by increasing their total cost. This final die test yield is close to 90% [33] and we assume it remains constant throughout the model's period of study.

Remark. In some literature [27, 42], the historical trend of the number of transistors has been used as a technical indicator of the evolution of the semiconductor industry and its ability to manufacture high density ICs. In this dissertation, the chief technical indicator is feature size. In retrospect, the feature size represents the ability of the manufacturer to produce high density ICs

and, in the chain of technology-driving trends, feature size drives the transistor densities to increase on a chip.

3.3 Magnetic Hard Disk Supply Model

In the early 1960s, digital magnetic recording was introduced in large computer systems because it could provide an efficient response to a data access, it was reliable, and it was low in cost [3, 57, 88]. IBM was the leading innovator in the technology [101] because it complemented its large-systems/large-database mainframe computers (which is still the bread and butter of IBM's business). As computers became smaller and the demand for them grew, the need for small, computer-resident, data storage devices emerged. Companies such as Conner Peripherals, IBM, Tandon, and others designed and produced several types of magnetic storage devices and supplemented them with portable and flexible magnetic diskettes. To differentiate the diskette from the computer-resident magnetic disk, the notion of magnetic hard²⁰ disk storage emerged. The data retained in magnetic storage is permanent because the magnets or the bit cells where the data is stored are permanent magnets.

The magnetic storage technology has been challenged by optical and magneto-optical storage technologies. However, with several improvements to the magnetic recording media, the read and write heads, the channel's electronic signal processing, and the packaging techniques, the magnetic technology has so far been able to maintain the lead among all the information recording sectors

²⁰In the literature, the words "hard" and "rigid" are synonymous when used in describing a magnetic disk storage device.

of the computer industry. For more details on the technological improvements, check the following references [3, 5, 57, 88, 101, 105].

Subsection 3.3.1 presents historical data on the physical characteristics trends of magnetic hard disks, Subsection 3.3.2 presents historical data on capabilities and price trends of rigid magnetic disks, Subsection 3.3.3 presents the assumptions and the terminology used in the development of the supply model of magnetic hard disks, and Subsections 3.3.4 through 3.3.11 present the mathematical formulation of the magnetic hard disk supply model.

3.3.1 Historical Data on the Physical Characteristics of Magnetic Hard Disks

Magnetic hard disk production is a worldwide activity, and several companies play major roles in moving a final product to market. Multiple steps are involved in the manufacturing of each hard disk subcomponent. The higher the accuracy of production is, the higher the reliability and performance of the disk.

As presented in Section 2.3, a magnetic storage system can have more than one disk, with read, write, and erase heads for each disk surface. The disk surface is a series of concentric circular tracks, where each track is formed by a concatenation of magnetic bit cells. The heads are moved by an actuator, controlled by a servomechanism which computes the distance the heads must move before reaching the required data. The data is received and transmitted by the hard disk through a channel. This channel is characterized by a maximum data rate at which the hard disk can operate and a microcontroller that coordinates

the data transfer.

For the purposes of reviewing historical trends in the physical characteristics of magnetic hard disks, the subcomponents of the magnetic storage devices are grouped in three categories. The first category is the disk and the magnetic medium, the second is the read/write/erase heads and the actuator, and the third is the data channel. Figure 3.10 presents the technology-driving trends [3] of the first and second categories in a log-linear graph, where the vertical axis indicates the base 10 logarithm of their values in nanometers and the horizontal axis indicates the year of the IBM's DASD²¹ market introduction. The IBM DASDs product numbers are listed directly above the horizontal axis. For instance, the products IBM 3380 and 3380E were introduced in 1982 and 1985, respectively.

The trends of each of the three categories of magnetic storage subcomponents listed above are discussed in the following paragraphs, as well as the behavior over time of the head-actuator setup width.

Disk and Magnetic Medium. The disk and magnetic medium technology-driving trends are the disk track pitch, the bit cell length, and the medium thickness. Keep in mind that as the track pitch decreases, the areal density of the disk increases and, as the bit cell length decreases, both the areal and the linear densities increase. The medium thickness measures the thickness of the magnet-sensitive film coating on the disk. Figure 3.10 shows that it exhibits a decreasing trend. This is due to the use of thinner films of magnetic materials

²¹DASD is equivalent to Direct Access Storage Device.

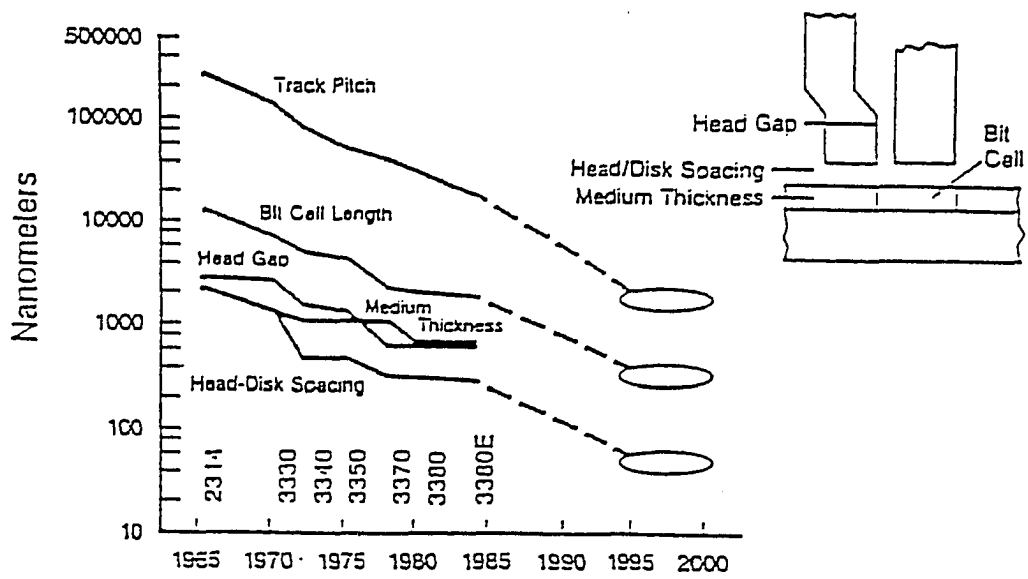


Figure 3.10: DASD recording system scaling. Source: [3].

in the production process. A thinner media has also improved the track pitch and the bit cells layout and increased the areal density of the disk. The trends of the track pitch, the bit cell length, and the medium thickness in Figure 3.10 can be expressed mathematically as follows:

$$TP_t = 310 * 10^{-0.066*(t-1965)} \text{ (micron)} \quad (3.37)$$

$$BCL_t = 13.5 * 10^{-0.049*(t-1965)} \text{ (micron)} \quad (3.38)$$

$$MT_t = 2.11 * 10^{-0.0295*(t-1965)} \text{ (micron)} \quad (3.39)$$

where TP_t represents the track pitch in microns, BCL_t represents the magnetic bit cell length in microns, MT_t represents the medium thickness in microns, and t represents the year in which the track pitch, the bit cell length, and the medium thickness are calculated.

Read/Write/Erase Heads and Actuator. The inductive read/write/erase head technology-driving trends are measured by the head gap spacing and the head-medium spacing. Illustrated in the top right of Figure 3.10, the head gap is shown to be the distance separating the poles of the magnetic core of the head, and the head-medium spacing is shown to be the distance at which the head flies above the disk surface. A smaller head gap complements a smaller bit cell length, improving both the linear and areal densities of the disk. A smaller head-medium spacing decreases the data misregistration errors, complements the rotational speed of the disk, and improves the data rates (megabytes/second) at which the disk can operate. The trends of the head gap spacing and the head-medium spacing are traced in Figure 3.10. They can be approximated

mathematically by Equations 3.40 and 3.41, respectively:

$$HGS_t = 2.8 * 10^{-0.04*(t-1965)} \text{ (micron)} \quad (3.40)$$

$$HMS_t = 2.11 * 10^{-0.055*(t-1965)} \text{ (micron)} \quad (3.41)$$

where HGS_t represents the head gap spacing in microns, HMS_t represents the head-medium spacing in microns, and t represents the year in which the head gap and the head-medium spacings are calculated.

No trends were available on actual performance measures of the servo and the actuator²², such as data-access times, seek times, and latency. Nevertheless, improvements have been made to integrate the head, the actuator, and the servo in one device by using a technique called micromechatronics [88], which offers improvements in the overall performance levels of the disk. The effect of this integration on the hard disk's attributes will be reflected through the cost per megabyte trend, to be presented later in the model formulation subsection. What follows is a description of the head-actuator setup factor and how it affects the magnetic hard disk's volumetric density.

Head-Actuator Setup. As the technology-driving trends in Figure 3.10 improve over time, the areal density and the data rate of the magnetic hard disk improve. Furthermore, the volumetric density of the disk is improved by the reduced head-medium spacing, the reduced medium thickness, and the overall smaller head-actuator setup [88]. A decrease in the head-actuator setup width increases

²²The servo and actuator performance measures depend on the configuration and the physical dimensions of the magnetic hard disk—all manufacturer dependent characteristics on which no trends were available.

the number of disks to be packaged in a fixed height storage system. As of 1991, the head-actuator setup width was in the 0.4 to 0.5 centimeter range. Unfortunately, no head-actuator setup width data for other years was available, so no trend could be estimated from historical data. We assumed the head-actuator setup width to be decreasing at approximately the same rate as the technology-driving trends provided in Figure 3.10, making it possible to express its behavior over time as follows:

$$HASW_t = HASW_0 * 10^{-0.04*(t-1991)} = 0.45 * 10^{-0.04*(t-1991)} \text{ (cm)} \quad (3.42)$$

where $HASW_t$ represents the head-actuator setup width in centimeters, $HASW_0$ represents the initial value of the head-actuator setup width—assumed to be 0.45 cm as of 1991—and 0.04 represents an average of the slopes of the lines in Figure 3.10. The t term represents the year in which the head-actuator setup width is calculated.

Data Channel. The hard disk's data channel performance is characterized by its data bandwidth, or data rate, measured in megabytes per second. A microcontroller is located at the end of the channel, decoding and queuing disk data-access commands. As the disk responds with the requested data, the microcontroller receives the data and feeds it to the channel, thereby controlling the data rate. The data rate of the channel must be at least equal to the data response rate of the disks. The technology-driving trends of the disk microcontroller are the same as those for microprocessors, so the MIPS rating of the microprocessors is used as a *relative* measure to reflect the performance of the data channel. Since most microcontrollers have been designed along the lines of

the CISC architecture [71], the CISC MIPS rating behavior of the ICs supply model will be used in the equation capturing the effects of the data channel on the overall supply of the magnetic hard disk.

3.3.2 Historical Data on Magnetic Hard Disk Capabilities and Price Trends

The following paragraphs present the price per megabyte and areal density trends of the magnetic hard disk storage. The price per megabyte encompasses the price contribution of all the hard disk's components, and the areal density encompasses the linear and track densities of the hard disk. The volumetric density of the disk depends on the head-actuator setup width, the trend of which, unfortunately, was not available to us²³ so that no historical trend on volumetric density could be estimated.

Magnetic Hard Disk Price/MB Trend. Two log-linear graphs are presented in Figure 3.11, the first being the price per megabyte of large capacity magnetic hard disks and the second the price per megabyte of DRAM chips [88]. The vertical axis indicates the base 10 logarithm of the prices in \$/MB and the horizontal axis indicates the time. The large capacity rigid disk part of Figure 3.11 has two lines, one indicating the custom price trend and another indicating the original equipment manufacturer (OEM) price trend. The two lines are parallel, and their values differ by almost 100%. The behavior of the actual price per megabyte trend of large capacity magnetic hard disks can be

²³In Equation 3.42, the behavior of the head-actuator setup width over time was assumed.

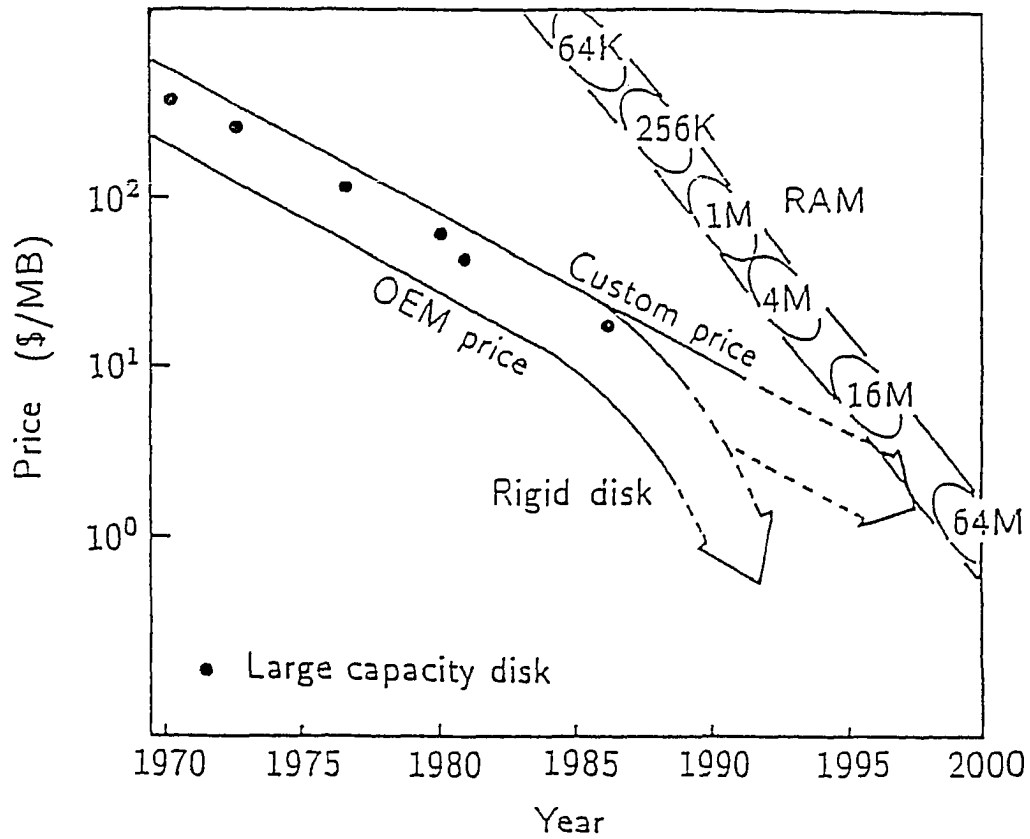


Figure 3.11: Actual price/megabyte of magnetic hard disk storage versus time.
Source: [88].

expressed mathematically as follows:

$$PMBD_t = 404.57 * 10^{-0.0857*(t-1970)} \quad (\$/MB) \quad (3.43)$$

where $PMBD_t$ represents the actual dollar price per megabyte, and t represents the year in which the actual price per megabyte is calculated.

If we assume that the manufacturer marks up the direct costs of the large capacity magnetic hard disks by 200% before selling them to the OEMs or retail companies, the cost per megabyte can be computed from Equation 3.43 by dividing the price by 3:

$$CMBD_t = 134.86 * 10^{-0.0857*(t-1970)} \quad (\$/MB). \quad (3.44)$$

Equation 3.44 is only a cost per megabyte estimation to be used in tuning the results of the magnetic storage model. The manufacturer's actual costs could be lower or higher, depending on the manufacturer's position on the learning curves and the attributes of the manufactured hard disks.

Hard Disk Areal Density Trend. The areal density trends of several magnetic storage devices are presented in Figure 3.12 as log-linear graphs, where the vertical axis indicates the magnetic device density in bits per mm^2 , and the horizontal axis indicates the product introduction year. Each device trend is labeled. For instance, the rigid disk's density trend is labeled RD, the flexible disk's FD, the home video tape's HV, the professional video tape's PV, the audio tape's AT, and the data tape's DT [57]. The rigid disk's areal density is the trend of interest and can be expressed mathematically as follows:

$$RDAD_t = 64.53 * 10^{0.138*(t-1965)} \quad (bits/mm^2) \quad (3.45)$$

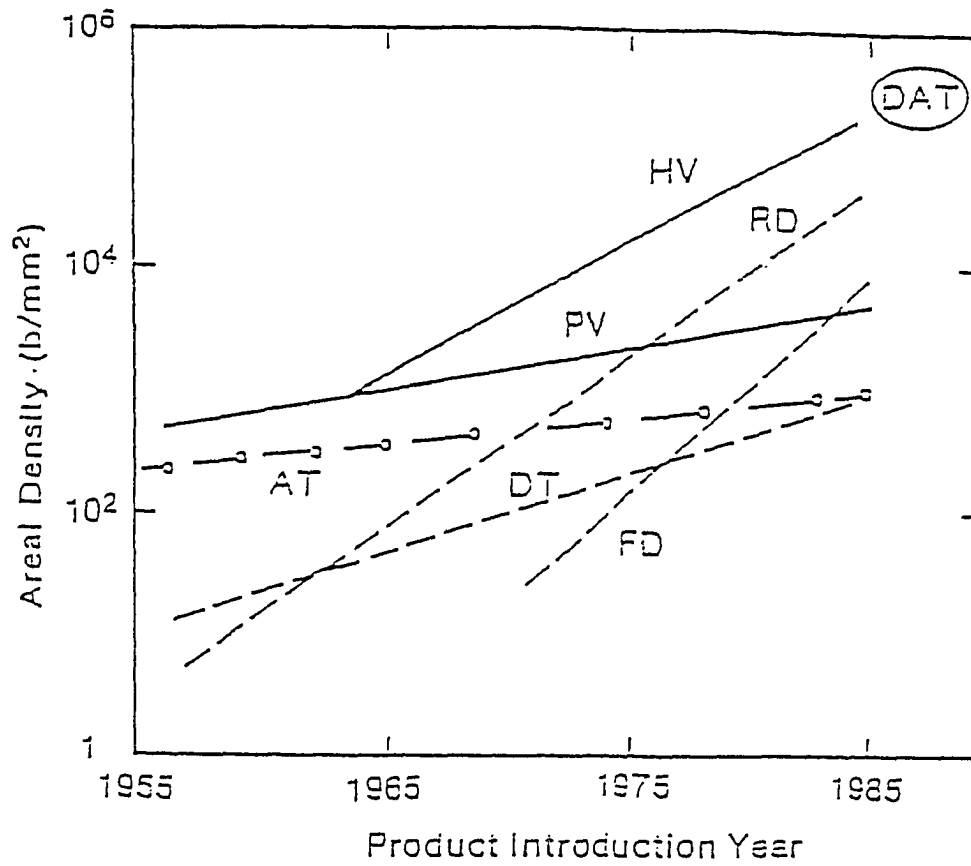


Figure 3.12: Areal densities of magnetic storage devices in bits/mm² versus time. Source: [57].

where $RDAD_t$ represents the rigid disk's areal density in bits per mm^2 , and t represents the year in which the areal density is calculated.

3.3.3 Model Assumptions and Terminology

The digital file storage supply model incorporates equations capturing the behaviors of magnetic hard disk attributes, specifically, the cost per megabyte, the data rate, and the areal density of rigid magnetic storage media. Before presenting the mathematical formulation of the model, several assumptions are listed. This is followed by a list defining the variables and the parameters used in the model.

Model Assumptions. In modeling the dynamics of the supply of magnetic hard disks, the following assumptions are made:

- Each disk is magnetically coated on both surfaces.
- Each disk surface has one single read/write/erase head.
- All the disks surfaces can be read or written at the same time.
- When the storage system is powered on, the disks reach a design specific rotational speed and maintain this speed throughout the system's operations.
- The storage system's microcontroller and channel have a constant operational speed and a constant bandwidth, respectively. Consequently, the system has a constant data rate.

- The storage system dimensional specifications, such as the height and the disk diameter, are assumed to remain constant over the period of study.
- Technological trends of the past in overcoming physical manufacturing barriers continue during the period of study.
- Past trends in magnetic hard disk manufacturing yields will continue to increase or, at worst, remain constant.
- Past trends in magnetic hard disk testing will continue to improve to deal with the higher density disks, thereby improving their reliability.

Definitions and Terminology. In presenting the model's equations, several parameters and variables are introduced, and they are defined as follows:

0	the time index for the first year of the study period
t	the time index
TP	the track pitch, in microns
BCL	the bit cell length, in microns
HGS	the head gap spacing, in microns
HMS	the head-medium spacing, in microns
MT	the magnetic medium thickness, in microns
DT	the metal disk thickness, in microns
$HASW$	the head-actuator setup width, in cm
$RSUP$	the hard disk support area radius, in cm
$RSTO$	the hard disk storage area radius, in cm
DD	the hard disk diameter, in cm

<i>DH</i>	the height of the hard disk box, in cm
<i>DPER</i>	the number of disks per centimeter of box height
<i>X</i>	the number of tracks on each disk surface
<i>RDAD</i>	the rigid disk areal density, in megabytes/cm ²
<i>TC</i>	the track capacity, in megabytes
<i>LD</i>	the linear density per track, in megabytes/cm
<i>TD</i>	the track density across the disk surface, in tracks/cm
<i>AC</i>	the areal capacity per disk surface, in megabytes
<i>AD</i>	the areal density per disk surface, in megabytes/cm ²
<i>VC</i>	the volumetric capacity of the storage system, in megabytes
<i>VD</i>	the volumetric density of the storage system, in megabytes/cm ³
<i>MIPS</i>	the number of million instructions per second
<i>DRPM</i>	the number of disk rotations per minute
<i>DR</i>	the storage system data rate, in megabytes/second
<i>CMB</i>	the cost per megabyte (model results), in dollars/megabyte
<i>CMBD</i>	the cost per megabyte (actual data), in dollars/megabyte
<i>TCD</i>	the total cost of the magnetic hard disk storage system, in dollars

The mathematical formulation of the magnetic hard disk model will proceed as follows: first, in Subsection 3.3.4 the radius of the magnetic storage media is computed; second, in Subsection 3.3.5 the number of disk recording tracks is calculated; third, in Subsection 3.3.6 the track capacity and density are computed; fourth, in Subsection 3.3.7 the data rate computation is presented, followed by a calculation of the optimal magnetic storage radius when the data rate is constant; fifth and sixth, in Subsections 3.3.8 and 3.3.9 the areal and volumetric capacities and densities are calculated, respectively; seventh, in Subsection 3.3.10 the

hard disk cost per megabyte is expressed as a function of several magnetic and semiconductor technology-driving trends; and eighth, in Subsection 3.3.11 the total hard disk cost is computed.

3.3.4 Magnetic Storage Radius

As shown earlier, a magnetic hard disk consists of a stack of disks, each separated from the other by a head-actuator setup width. The disks are coated on each surface with magnetic media capable of permanently storing the digital data in a format of polarized magnets. Each disk surface is divided into X number of tracks, separated from each other by $X - 1$ number of track pitches. However, part of the disk surface is used as a rotational support, so the width of the magnetic storage surface, or what we refer to as the magnetic storage radius, is equal to:

$$RSTO = \frac{DD}{2} - RSUP \quad (cm) \quad (3.46)$$

where $\frac{DD}{2}$ is equivalent to the disk radius and $RSUP$ is the disk support radius. See Figure 3.13 for an illustration of the disk parameters used in Equation 3.46.

3.3.5 Number of Disk Recording Tracks

We assume that the width of a track is equal to one bit cell length, and relating the number of tracks, the number of track pitches, and the storage radius, the following equation is obtained:

$$BCL_t * X + TP_t * (X - 1) = RSTO * 10^4. \quad (3.47)$$

The storage radius ($RSTO$) is actually the distance between the support radius ($RSUP$) and the rim of the disk. We refer to it as a storage “radius” in order to

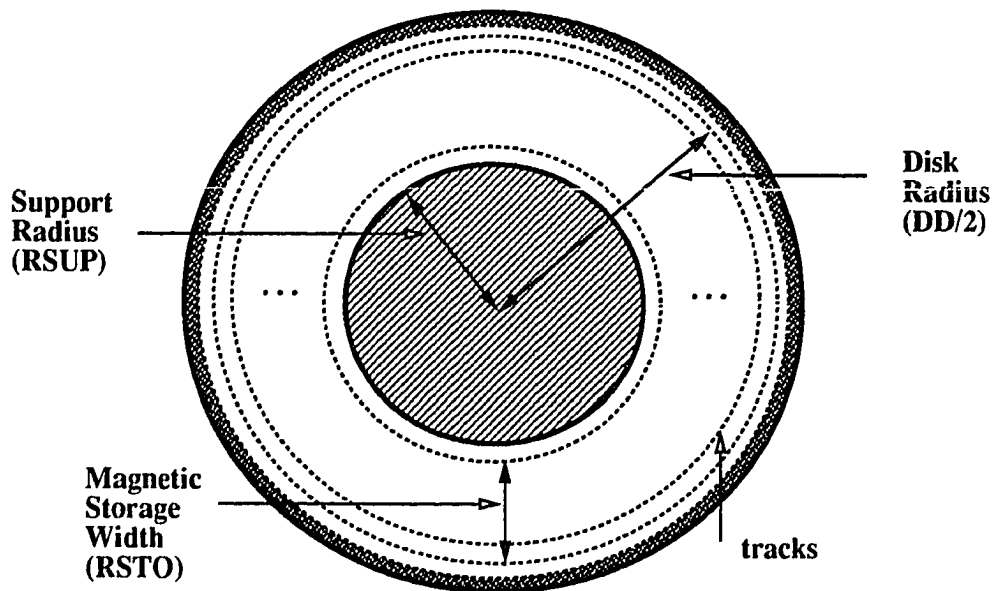


Figure 3.13: Illustration of the dimensional parameters of a magnetic hard disk.

be consistent with the fact that this distance is filled with circular tracks. From Equation 3.47, the number of tracks X is computed as:

$$X = \frac{RSTO * 10^4 + TP_t}{BCL_t + TP_t} \approx \frac{RSTO * 10^4}{TP_t} \quad (3.48)$$

where X is approximated as the ratio of the storage radius and the track pitch at t because the value of the bit cell length is negligible when compared to the track pitch's.

3.3.6 Track Capacity and Density

Each track has the same number of sectors, and, since the disk is rotating at a constant speed and has a constant data rate, each sector—and, consequently, each track—has the same number of bits [57]. Since the support radius ($RSUP$) has the smallest track in circumference and all the tracks have the same capacity, the number of bits on this track sets the standard number of bits that each track can have. The track capacity in megabytes²⁴ is:

$$TC_t = \frac{2 * \pi * RSUP * 10^4}{BCL_t} \text{ (bits)} = \frac{2 * \pi * RSUP * 10^4}{BCL_t * (8 * 10^6)} \text{ (MB)}. \quad (3.49)$$

The linear bit density per track is equal to the number of bit cells in one centimeter, and its equation is:

$$LD_t = \frac{10^4}{BCL_t * (8 * 10^6)} \text{ (MB/cm)} \quad (3.50)$$

and the track density across the storage surface is:

$$TD_t = \frac{10^4}{TP_t} \text{ (tracks/cm)}. \quad (3.51)$$

²⁴Since there are 8 bits to a byte, $8 * 10^6$ is the product used to transform the track capacity values from bits to megabytes.

3.3.7 Hard Disk Data Rate

Before computing the areal and volumetric capacities of the storage system, since the data rate has been mentioned several times, it is appropriate at this point to present its formulation. What follows is the mathematical expression of the data rate, with a description of each of the expression's factors. The data rate equation is:

$$DR_t = TC_t * \frac{DRPM_t}{60} * [(DPER_t * DH)] \quad (MB/sec). \quad (3.52)$$

The first term in the data rate equation is the **track capacity (TC)** in megabytes, the formulation of which is given by Equation 3.49. The second term is the **number of disk rotations per minute (DRPM)**, expressed in seconds by dividing it by 60. The DRPM depends on the rotating motor speed, on the channel's bandwidth, and on the head and media characteristics handling the read and write accesses. For the model, the DRPM trend was assumed to increase at 100 rotations per year since 1980, with an initial value of 3000. As of 1991, the DRPM was in the 3600-4200 range and, based on conversations with colleagues, the number of disk rotations per minute was expected to reach 5000 by the year 2000. The DRPM equation over time is:

$$DRPM_t = 3000 + (t - 1980) * 100 \quad (rot/min) \quad (3.53)$$

where t represents the year in which the DRPM is calculated. The DRPM will saturate over time due to the high heat generated from the rotations and the high power it absorbs to reach them [57]. In 1991, an average value of DRPM is 4100 rotations per minute. (Most manufacturers now rely on techniques other than increased rotational speed to improve their data rates.)

The third term in Equation 3.52 captures the **number of disk surfaces** off which data can be read or onto which data can be written. For a given storage system box height, the number of disks that can fit depends on the head-actuator setup width, the disk thickness, the medium thickness, and the head-medium spacing. Since each disk surface has one read/write/erase head and since all the surfaces can be read or written at the same time, the actual data rate of the system is equal to the data rate per one disk surface, which is equal to the product of the first and the second terms of Equation 3.52, multiplied by the number of surfaces written or read. Because the disks are coated on each side, each disk has two heads associated with it, two head-actuator setups, two media surfaces, two head-medium spacings, and one disk thickness; the equation for the number of disks per centimeter is:

$$DPER_t = \frac{1}{[2*(MT_t + HMS_t) + DT_t] * 10^{-4} + 2*HASW_t} \approx \frac{1}{2*HASW_t} \quad (disks/cm). \quad (3.54)$$

Historically, the number of disks per centimeter (DPER) was strongly influenced by improvements to the head-actuator setup width (HASW), the assumed trend of which was given as Equation 3.42. In Equation 3.52, the product of the storage system's height (DH), in centimeters, and the number of disks per centimeter (DPER) is equal to the total number of disks assembled in the storage system.

The data rate of a storage system is not an attribute which could easily have a general behavioral trend over time. This is due to the data rate's dependency on the configuration and physical dimensions of the storage system, both of which do not follow a standard dimension scheme. Nevertheless, the data rate trend is on the increase.

Optimal Hard Disk Storage Radius. For a storage system with a constant data rate, it is demonstrated in [57] that the track capacity and, consequently, the storage capacity are maximized by having the support radius, $RSUP$, equal to half the disk radius:

$$RSUP_{opt} = \frac{DD}{4} \text{ (cm)}. \quad (3.55)$$

From Equations 3.46 and 3.55, the optimal width of the storage surface, or what we refer to as the optimal storage radius, is equal to the support radius:

$$RSTO_{opt} = RSUP_{opt} = \frac{DD}{4} \text{ (cm)}. \quad (3.56)$$

3.3.8 Areal Capacity and Density

As shown earlier, all the tracks on a disk surface have the same capacity. So the areal capacity per disk surface is simply equal to the product of the track capacity and the number of tracks:

$$AC_t = TC_t * X \text{ (MB)}. \quad (3.57)$$

From Equations 3.50 and 3.51, the areal density can be computed as the product of the linear bit density and the track density:

$$AD_t = LD_t * TD_t \text{ (MB/cm}^2\text{)}. \quad (3.58)$$

3.3.9 Volumetric Capacity and Density

The volumetric capacity of the storage system is equal to the number of disks in the system multiplied by twice the areal capacity per disk surface, since the disks are coated on both surfaces:

$$VC_t = 2 * AC_t * [(DPER_t * DH)] \text{ (MB)} \quad (3.59)$$

and the volumetric density is equal to the product of twice the disk areal density and the number of disks per centimeter:

$$VD_t = 2 * AD_t * [DPER_t] \quad (MB/cm^3). \quad (3.60)$$

3.3.10 Magnetic Hard Disk Cost/MB

The final equations of the model compute the cost per megabyte of magnetic hard disk storage and, ultimately, the total cost of the storage system. As discussed earlier, the storage system has three groupings of subcomponents, and each of these subcomponents has a set of technology-driving trends. Figure 3.14 illustrates, in a relational diagram, the effect of each subcomponent's set of technology-driving trends on the overall cost per megabyte of magnetic storage:

- Head gap spacing and head-medium spacing were chosen as magnetic hard disk head technology-driving trends and, indirectly, magnetic hard disk cost-driving trends, because achieving lower spacings requires much scientific research, effort, and R&D money, all of which increase the magnetic hard disk cost per megabyte.
- Bit cell length, medium thickness, and track pitch were chosen as disk technology-driving trends and, indirectly, magnetic hard disk cost-driving trends, because a smaller bit cell length is harder to deposit on the disk surface and needs fancier head and servomechanism designs to handle the higher bit densities resulting from a smaller magnetic cell length; similarly, it is scientifically challenging to achieve lower track pitches and lower medi-

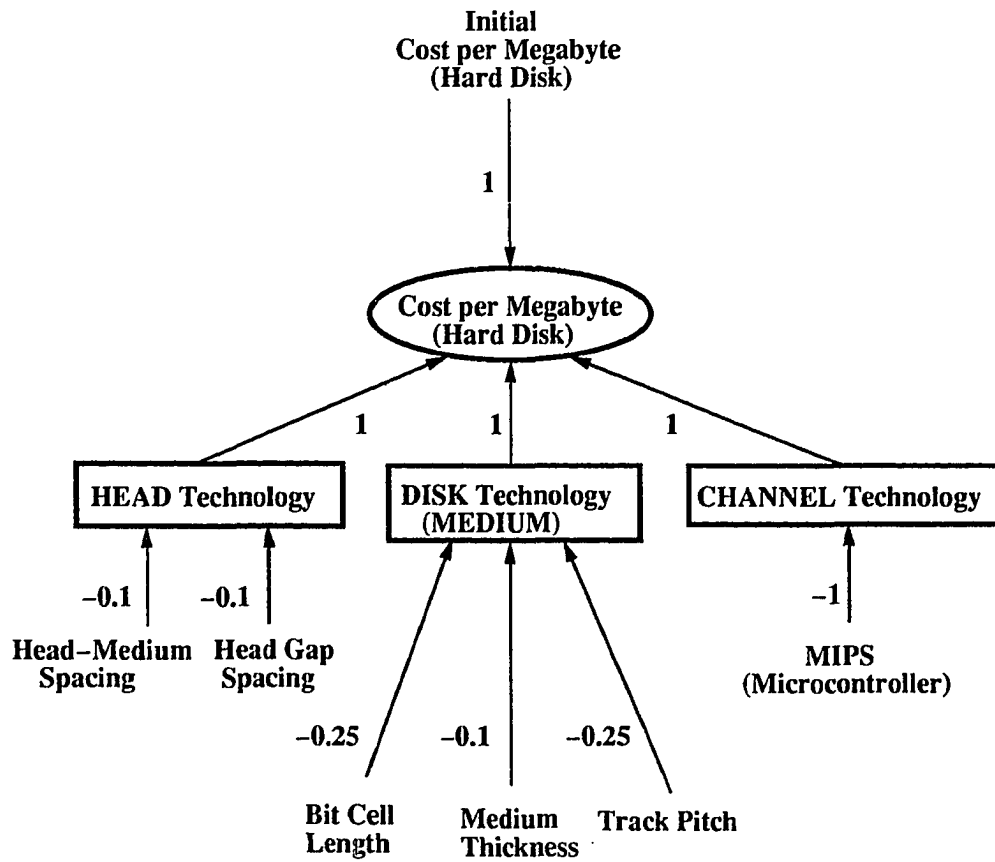


Figure 3.14: Relational diagram of the cost/megabyte of a magnetic hard disk.

um thicknesses, all costly R&D outlays for disk production, and factors which increase the magnetic hard disk cost per megabyte.

- MIPS was chosen as a channel technology-driving trend and, indirectly, a magnetic hard disk cost-driving trend, because as the MIPS rating of the channel's microcontroller improves, the price per MIPS decreases (see actual data in Section 3.2 and the ICs supply model results in Section 4.3). Thus, if the data rate of the disk is assumed to remain the same, the cost of the channel decreases and, consequently, the cost per megabyte of magnetic hard disk.

Expressed mathematically, the cost per megabyte is equal to:

$$CMB_t = CMB_0 * HEAD_t * DISK_t * CHANNEL_t \quad (\$/MB) \quad (3.61)$$

where

$$HEAD_t = \left(\frac{HGS_t}{HGS_0}\right)^{-0.1} * \left(\frac{HMS_t}{HMS_0}\right)^{-0.1} \quad (3.62)$$

$$DISK_t = \left(\frac{BCL_t}{BCL_0}\right)^{-0.25} * \left(\frac{MT_t}{MT_0}\right)^{-0.1} * \left(\frac{TP_t}{TP_0}\right)^{-0.25} \quad (3.63)$$

$$CHANNEL_t = \left(\frac{MIPS_t}{MIPS_0}\right)^{-1} \quad (3.64)$$

and the initial cost per megabyte, CMB_0 , is computed from Equation 3.44 for the initial year in the period of study. (It is assumed that CMB_0 includes the cost of labor and raw materials, in particular the actuator and the servomechanism costs mentioned earlier.) Equations 3.62 and 3.63 reflect how the cost per megabyte of magnetic hard disk storage in Equation 3.61 increases as the head gap spacing, the head-medium spacing, the bit cell length, the medium thickness, and the track pitch decrease. Equation 3.64 reflects how an increase in the MIPS rating

of the storage system's microcontroller decreases the actual cost per MIPS of the chip and, consequently, the cost per megabyte of magnetic storage. The values of the exponents, (a_n 's), in Equations 3.62, 3.63, and 3.64 were obtained by tuning the model to yield results to match the actual cost per megabyte data of hard disk storage provided in Subsection 3.3.2. All of the $|a_n|$'s are ≤ 1 , and each reflects an estimate of its relative importance to the behavior of the cost per megabyte over time. For example, a 10% decrease in the bit cell length leads to an approximate increase of 2.5% in the cost per megabyte (see Equation 3.9), and similarly for other factors of Equations 3.62, 3.63, and 3.64.

3.3.11 Magnetic Hard Disk Total Cost

Finally, the total cost of the magnetic hard disk storage system is equal to the product of the volumetric capacity and the cost per megabyte:

$$TCD_t = VC_t * CMB_t \text{ (\$)}. \quad (3.65)$$

3.4 Color CRT Display Supply Model

The CRT is still, after 100 years of experience in its manufacturing, a large, heavy, and power-hungry display technology. Its popularity can be attributed to its diversified range of applications, its fast response time, and its low cost relative to the other display technologies [90]. As of 1991, the CRT has the highest market share of displays sold [91], and users continue to appeal to the monitors' manufacturers to reduce their prices, increase the CRT's resolution, and eliminate the glare [2]. Other than in home televisions, the CRT capabilities are most apparent in the computer workstation displays, where CAD

and computer simulations and animations are the most prevalent applications and where the CRT's speed of response in an interactive environment is key.

The size and power consumption of the CRT are a consequence of the physics of its design. These disadvantages have been exploited by leaner battery-operated displays, which may overtake the CRT's market share lead, at least in the computer market. The liquid crystal based display (LCD) is the leading challenger and has the highest potential of displacing the CRT's market share. Already 10-inch color LCDs are available in portable computers [69], and before long workstation LCDs will be developed and sold as independently packaged computer displays, as CRTs are now.

Most workstation CRT displays measure 19 to 20 inches on the diagonal, and the screens are available in color or B&W. The CRT supply model, presented later in this section, captures the cost per megapixel trend of a 19-inch color CRT and relates it to the resolution technology-driving trends such as the metal shadow mask hole pitch, the speed of the screen-driving circuitry, and the enhancements to the electron beams generation, acceleration, and deflection hardware used. Before presenting the mathematical formulation of the color CRT display supply model, historical data on the color CRT's physical characteristics are presented in Subsection 3.4.1, and historical data on the color CRT's capabilities and price trends are presented in Subsection 3.4.2.

3.4.1 Historical Data on the Physical Characteristics of Color CRT Displays

As discussed in Section 2.4, the sharpness and the resolution of a CRT picture are inversely proportional to the size of the pixel or the size of the smallest picture element that can be displayed within the limitations and parameters of the display structure and hardware. In workstations today, 1+ megapixel color CRT displays are very common. Display hardware drivers can be expensive at times, depending on the application and the color palettes desired [33].

In response to the increased demand for higher resolutions and more colors per pixel, CRT production has been subjected to several enhancements and design changes over the years. Recent VLSI improvements have shifted the focus for improved resolution from the CRT driving circuitry to the electron beams generation, focus, and deflection hardware, and to the manufacture of the tube, including the metal shadow mask used with color CRTs [8, 14].

When excluding the driving circuitry, the main parts of the color CRT structure are the tube, the electron guns, the electron beams' accelerating, focusing, and deflecting apertures, and the metal shadow mask.

Since the 19-inch color CRT workstation monitor was only introduced in 1985 [16, 33], it has a brief history. The current leading manufacturers are Japanese, companies like Hitachi, Mitsubishi, NEC, and Sony, whose engineering facilities are spread between the far eastern rim of Asia and the western United States, so obtaining any manufacturing trends has been very difficult. Several companies, like Stanford Resources and Tannas Electronics, have compiled reports and books on the CRT production trends, but they are costly and largely

limited to future projections based on past behavior, rather than historical data. The only actual physical characteristic data found tracks the hole pitch trend of the metal shadow mask.

Metal Shadow Mask Hole Pitch. The metal shadow mask hole pitch data is listed in Column 5 of Table 3.7, with values dating back to 1982. If one traces a curve to the hole pitch's trend over time, the result can be expressed mathematically as follows:

$$HP_t = 0.6 * 10^{-0.04*(t-1982)} \text{ (mm)} \quad (3.66)$$

where HP_t represents the hole pitch in millimeters, and t represents the year in which the hole pitch is calculated. The number of holes per inch (HPI_t), or the maximum number of pixels per inch ($MaxPPI_t$), is equal to:

$$HPI_t = MaxPPI_t = \frac{25.4}{HP_t} \text{ (#holes/in or Max #pixels/inch)} \quad (3.67)$$

where 25.4 is equal to the number of millimeters per inch.

It was stated earlier that improved ICs manufacturing capabilities will lead to faster and denser microcontrollers and DRAMs, respectively, and, therefore, increase the number of holes per inch. Another important technology which enables a continuous increase in the number of holes per inch—and, consequently, the CRT's resolution—is lithography. Lithography is used to deposit the color phosphors on the inner face plate of the bulb, one color dot at a time. The more sophisticated lithography is, the more in pace the phosphors resolution on the face plate will be with the holes resolution of the metal shadow mask. Unfortunately, very small hole pitches reduce the metal shadow mask production

yields and, consequently, increase the display costs, thus making the monitors prohibitively expensive for general workstation users.

3.4.2 Historical Data on Color CRT Display Capabilities and Price Trends

Capabilities and price data of the 19- and 20-inch color computer monitors is presented on Table 3.7 (unavailable data is listed as n.a.). Under Column 2, the CRT manufacturers are listed. Sony is among the listed manufacturers; however, Sony uses a different shadow mask technology than that considered in the CRT model presented in this section. Still, Sony's monitor resolutions and prices reflect the general trends of the industry. Their color CRTs use the Trinitron specifications: each of the three color phosphors—red, green, or blue—are deposited in vertical stripes along the inner screen of the tube, and each stripe has a different color than those adjacent to it. The deposition pattern is repeated along the screen, and three electron guns, one per color, are used to excite the phosphors, simultaneously, while scanning the screen and generating the image. The shadow mask could be considered as a striped metal sheet, located behind the screen, where each stripe corresponds to a particular color phosphor. The mask technology considered in this section's CRT supply model uses a perforated metal shadow mask, where each perforated hole corresponds to three phosphor dots, each with a different color. The three electron beams must go through the mask hole and hit the dots, simultaneously, for the corresponding color light to be emitted.

Most of the monitors listed on Table 3.7 measure 19 inches on the

Color CRTs							
1	2	3	4	5	6	7	8
Year	Name Brand	Diagonal Size (inch)	Resolution (HxV) or (MP)	Hole Pitch (mm)	Refresh Rate (Hz)	Bandwidth (MHz)	Price (Actual Data) (\$)
1982	n.a.	19	n.a.	0.6	n.a.	n.a.	n.a.
1985	Sony	19	1	n.a.	n.a.	n.a.	8K
1986	Sony	19	1	n.a.	n.a.	n.a.	6K
1987	Mitsubishi	20	1280x1024	0.31	50-75	100	3.8K
1988	n.a.	19	1024x1024	n.a.	n.a.	n.a.	2.5K-4K
1989	Sony	19	1	n.a.	n.a.	n.a.	5K
1989	Sony	19	2048x1536	0.19-0.2	80	262	n.a.
1989	NEC	19	1024x768	0.31	56-80	65	3.2K
1989	n.a.	n.a.	1024x768	n.a.	n.a.	n.a.	3K
1989	n.a.	n.a.	1600x1280	n.a.	n.a.	n.a.	5K-10K
1991	Sony	19	1280x1024	0.28	n.a.	n.a.	3K

Table 3.7: Actual market data of 19- and 20-inch color CRT displays. Sources: [1, 2, 16, 28, 38, 59, 75, 81, 83, 93].

diagonal. Their **resolutions** are listed in Column 4 and all exhibit the 4:3 aspect ratio, except for one, which can be verified by examining the number of horizontal pixels times the number of vertical pixels (HxV) displayed. Note that the **number of pixels** is either smaller or equal to the total number of perforated holes in the mask. The total number of pixels depends on the screen-driving circuitry and on the ability of the apertures to focus and deflect the beams correctly at very high bandwidths.

The **refresh rates** of the CRTs are listed in Column 6 and range from 50 to 80 image scans per second, mostly non-interlaced. Their values are mainly flicker dependent [90].

In Column 7, the **bandwidths** of the CRTs reflect how fast the electron guns must switch per second to scan and refresh the screen to avoid flicker. The values are a function of the resolution and the refresh rates used in the CRT design.

Finally, the **price** data of the monitors is listed in Column 8 of Table 3.7. These are single unit prices, and bulk²⁵ prices are often only half as much [16].

Price per Megapixel. A price per megapixel trend is provided in Column 2 of Table 3.8. It is computed by dividing the prices of the monitors on Table 3.7 by their corresponding resolutions. If an exponential curve is fitted to the data to capture the behavior of the price per megapixel over time, it can be expressed

²⁵Bulk is equivalent to 10,000+ monitors.

Color CRTs		
1	2	3
Year	Price/MP (Actual Data) (\$)	#Pixels/inch
1985	8000	76
1986	6000	76
1987	2900	84
1988	2380-3800	67
1989	2400-5000	76-135
1991	2300	84

Table 3.8: Actual price/megapixel and number of pixels/inch market data of 19- and 20-inch color CRT displays. Sources: [1, 2, 16, 28, 38, 59, 75, 81, 83, 93].

as follows:

$$PMPD_t = 8000 * 10^{-0.085*(t-1985)} \quad (\$/MP) \quad (3.68)$$

where $PMPD_t$ represents the actual dollar price per megapixel, and t represents the year in which the actual price per megapixel is calculated. Assuming a 200% cost markup was used by the manufacturers [33] in setting the prices on Table 3.7, the cost per megapixel trend can be obtained by dividing each price per megapixel by 3:

$$CMPD_t = 2700 * 10^{-0.085*(t-1985)} \quad (\$/MP). \quad (3.69)$$

Number of Pixels per Inch. The number of pixels per inch (PPID) values of the CRTs on Table 3.7 are listed in Column 3 of Table 3.8. $PPID_t$ can be computed from the following resolution equation:

$$RES_t = \frac{(HS_{in} * PPID_t) * (VS_{in} * PPID_t)}{10^6} \quad (megapixel) \quad (3.70)$$

where RES_t represents the display's resolution in megapixels. When factored out of Equation 3.70, $PPID_t$ can be expressed as:

$$PPID_t = \sqrt[2]{\frac{RES_t * 10^6}{HS_{in} * VS_{in}}} \quad (\#pixels/inch) \quad (3.71)$$

where the horizontal size (HS) and the vertical size (VS) of the CRT can be computed from the diagonal size (DS) by using the 4:3 aspect ratio:

$$HS_{in} = \frac{4}{5} * DS_{in} \quad (inch) \quad (3.72)$$

$$VS_{in} = \frac{3}{4} * HS_{in} = \frac{3}{5} * DS_{in} \quad (inch). \quad (3.73)$$

(The 10^6 factor is used to convert the resolution values from pixel to megapixel, and vice versa.)

3.4.3 Model Assumptions and Terminology

The supply model for computer displays incorporates equations capturing the attributes of 19-inch color CRT monitors, in particular, the cost per megapixel, the bandwidth, and the number of pixels per inch and its effect on the display's resolution. Before presenting the mathematical formulation of the model, several assumptions are listed. This is followed by a list defining the variables and the parameters used in the model.

Model Assumptions. In modeling the dynamics of the supply of 19-inch CRT displays, the following assumptions are used:

- The display's dimensional specifications, like the diagonal size and weight, are assumed to remain constant over the period of study. The diagonal size is measured in inches.
- Technological trends of the past in overcoming physical manufacturing barriers continue during the period of study.
- Past trends in color CRT display manufacturing yields will continue to increase or, at worst, remain constant.
- Past trends in color CRT display testing will continue to improve to deal with the higher resolution monitors and the higher power consumption associated with them, thereby improving the reliability of the CRTs.

Definitions and Terminology. In presenting the model's equations, several parameters and variables are introduced, and they are defined as follows:

<i>0</i>	the time index for the first year of the study period
<i>t</i>	the time index
<i>DS</i>	the diagonal size of the CRT, in inches
<i>HS</i>	the horizontal size of the CRT, in inches
<i>VS</i>	the vertical size of the CRT, in inches
<i>HP</i>	the hole pitch of the metal shadow mask, in millimeters
<i>HPI</i>	the number of holes per inch perforated in the metal shadow mask
<i>PPID</i>	the number of pixels per inch (actual data)
<i>PPI</i>	the number of pixels per inch (model results)
<i>MaxPPI</i>	the maximum number of pixels per inch
<i>RES</i>	the resolution of the display, in megapixels
<i>MaxRES</i>	the maximum resolution of the display, in megapixels
<i>MDPUL</i>	the number of metal shadow mask defects per unit length
<i>MY</i>	the metal shadow mask production yield, in %
<i>RR</i>	the screen refresh rate, in Hertz
<i>DB</i>	the display bandwidth, in megaHertz
<i>MIPS</i>	the number of million instructions per second
<i>MEMORY</i>	the capacity of a DRAM, in megabytes
<i>PMPD</i>	the price per megapixel (actual data), in dollars/megapixel
<i>CMPD</i>	the cost per megapixel (actual data), in dollars/megapixel
<i>CMP</i>	the cost per megapixel (model results), in dollars/megapixel
<i>TCM</i>	the total cost of the CRT monitor, in dollars

The mathematical formulation of the color CRT display supply model will proceed as follows: first, in Subsection 3.4.4 the bandwidth of the CRT is computed; second, in Subsection 3.4.5 the resolution's mathematical expression and

technology-driving trends are discussed; third, in Subsection 3.4.6 the metal shadow mask production yield is calculated; and fourth, in Subsection 3.4.7 the cost of the CRT and the cost per megapixel are computed, including descriptions of the technology-driving trends affecting their behavior over time.

3.4.4 Bandwidth

Bandwidth has been a key factor in the CRT's response time and picture stability. It is a function of the resolution and the refresh rate of the monitor, factors depending upon the pixel density on the screen. Bandwidth is measured in megaHertz (MHz), and it is equivalent to the electron beams' switching frequency while the screen image is scanned and refreshed. The display bandwidth (DB) equation can be expressed mathematically as follows:

$$DB_t = RES_t * RR \text{ (MHz)} \quad (3.74)$$

where the refresh rate (RR) is set to values so as to avoid flicker as the number of pixels increase and the colors become more defined [90].

3.4.5 Resolution

The display's resolution, given in Equation 3.70, has almost doubled since the introduction of the first color 19-inch CRT monitor in 1985 [16, 76]. Resolution is influenced by several technology-driving trends, the most important of which are the hole pitch of the metal shadow mask and the technology-driving trends of the screen hardware drivers.

Resolution and Metal Shadow Mask Hole Pitch. The smaller the hole pitch, the higher the resolution. For a particular hole pitch, the maximum attainable resolution can be obtained by replacing the maximum number of pixels per inch in Equation 3.67 in the resolution Equation 3.70:

$$MaxRES_t = \frac{(HS_{in} * MaxPPI_t) * (VS_{in} * MaxPPI_t)}{10^6} \text{ (megapixel)}. \quad (3.75)$$

However, the maximum resolution is rarely the manufacturing standard because of difficulties encountered in deflecting the electron beams to penetrate the exact assigned hole in the metal shadow mask to hit the corresponding color phosphors. So the actual number of pixels per inch is slightly lower than the number of holes per inch in the metal shadow mask.

Resolution and Screen Hardware Drivers. Figure 3.15 illustrates, in a relational diagram, what effects the screen hardware drivers of the CRT have on the number of pixels per inch and the resolution, namely the speed of the graphics microprocessors (and/or microcontrollers) used and the capacity of the DRAMs installed. The graphics microprocessors execute a series of instructions fed to them by the running application program. In turn, they access the display allocated DRAMs and send the fetched data through a chip that converts the digital signals to analog signals²⁶. The levels of the analog signals set the intensities of the electron beams and the color shades of the pixels. The speed at which the analog signals are fed to the CRT sets the bandwidth. It is apparent that the attributes of the hardware drivers of the CRT have been pivotal in obtaining

²⁶The chip that converts the digital signals to analog is referred to as a D/A converter.

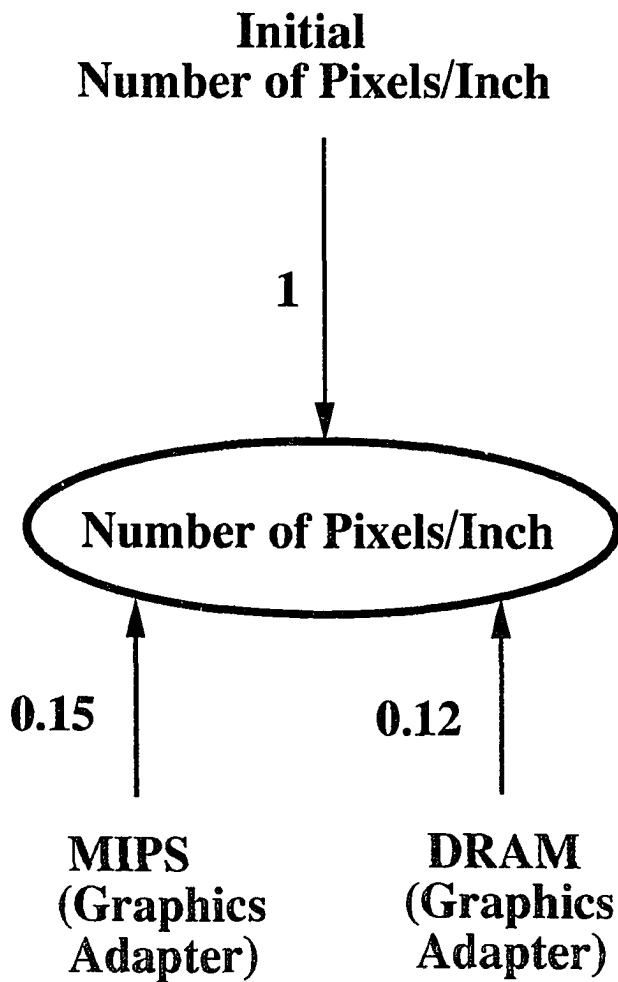


Figure 3.15: Relational diagram of the number of pixels/inch of a color CRT display.

higher resolutions and a larger number of colors and, in turn, in making the ICs technology-driving trends the CRT resolution's.

The top-of-the-line hardware drivers map each pixel to 24 bits of digital data, or 8 bits per color, and, since the digital bits are mapped onto analog signal levels, the actual number of colors that the screen can emit is equal to $2^{3*\#bits/color}$, in this case, 2^{24} or close to 17 million different colors [81].

The hardware drivers are referred to in the literature as graphics or video adapters. Most of the graphics microprocessors have been CISC architecture based [71]. Their relative performance can be measured in MIPS because the instructions are fairly simple and do not involve double precision floating point operations. The higher the MIPS rating, the faster the microprocessor executes the graphics instructions, and the faster the bits of each pixel get fed to the D/A converter to be displayed on the screen. Furthermore, as indicated in Figure 3.15, a higher DRAM capacity results in a cheaper cost per megabyte of DRAM, an increase in the number of DRAM chips that can be packaged on the graphics adapter, and a higher number of pixels per inch.

The effects of the accelerating, focusing, and deflecting apertures are not considered as factors of the number of pixels per inch behavior because their effects are cumulative, and there are no individual data trends relating them to the number of pixels per inch of color display. The cost per megapixel equation, to be presented later, will incorporate their effects as cost decreasing factors in an exponentially decreasing component of the equation.

The number of pixels per inch model, illustrated in Figure 3.15, can

be expressed mathematically as follows:

$$PPI_t = PPI_0 * \left(\frac{MIPS_t}{MIPS_0}\right)^{0.15} * \left(\frac{MEMORY_t}{MEMORY_0}\right)^{0.12} \quad (\#pixels/inch) \quad (3.76)$$

where PPI_0 corresponds to the initial value of the number of pixels per inch. The values of the exponents, (a_n s), in Equation 3.76 were obtained by tuning the model to yield results to match the actual number of pixels per inch data provided on Table 3.8. All of the $|a_n|$ s are ≤ 1 , and each reflects an estimate of its relative importance to the behavior of the number of pixels per inch over time. For example, a 10% increase in the microprocessor's MIPS rating, which might incorporate an increase in its speed and a more efficient instruction set (see Equation 3.20), is coupled with an approximate increase of 1.5% in the number of pixels per inch (see Equation 3.9), because faster adapters can drive a larger number of screen pixels. Also, a 10% increase in the capacity of the adapter's DRAM is coupled with an approximate increase of 1.2% in the number of pixels per inch, because more DRAMs enable more pixels to be stored during the refresh cycle.

3.4.6 Metal Shadow Mask Manufacturing Yield

As the hole pitch decreases, manufacturing the metal shadow mask becomes more intricate and more costly. Associated with the manufacturing of the shadow masks are their yields, which are directly proportional to the hole pitch values. Data on shadow mask production yields and mask defects per unit length are classified company information [16], so their actual values can only be estimated in their effects on the behavior of the cost per megapixel over time. In Equations 3.77 and 3.78, these effects are captured in the magnitude and sign

of the exponents. What follows is the mask's yield equation, which relates the number of mask production defects per unit length (MDPUL) to the diagonal size (DS) of the CRT:

$$MY_t = (MDPUL_t * DS_{in})^{\frac{-MDPUL_t}{4}} \quad (3.77)$$

where $MDPUL_t$ is expressed as:

$$MDPUL_t = MDPUL_0 * \left(\frac{HP_t}{HP_0}\right)^{-0.2} \quad (defects/inch) \quad (3.78)$$

where, as the hole pitch (HP) decreases, the number of mask defects increases and, consequently, the mask yield decreases.

3.4.7 Color CRT Display Cost

Now let's turn our attention to modeling the cost of the color CRT display. Since 1989, the prices of high resolution color CRTs decreased by approximately 10% per year [2], spurred by an increase in the number of manufacturers, enhanced production techniques, and higher CRT production yields. Any time a higher resolution is sought, the entire system's specifications change. The ability to transfer existing technology to new designs gives the CRT based displays an edge over other contemporary technologies. The total cost of a 19-inch color display is equal to the product of the resolution and the cost per megapixel:

$$TCM_t = RES_t * CMP_t \quad (\$). \quad (3.79)$$

The factors affecting the cost per megapixel of a color CRT display and the mathematical expression of the cost per megapixel are discussed below.

Cost per Megapixel. The overall cost per megapixel is a function of the quality of the display, its attributes, the precision of the design, the types of color phosphors used, the maximum bandwidth²⁷ it can handle, the hole pitch and the metal shadow mask yield, the graphics adapter, and the accelerating, focusing, and deflecting apertures. It is difficult to find a consistent display pricing strategy on the market because of the previously listed factors.

Figure 3.16 illustrates, in a relational diagram, the effects on the overall CRT cost per megapixel of the metal shadow mask production yields, the MIPS rating of the graphics adapter, and the amount of DRAM installed:

- Metal shadow mask yield was chosen as a CRT technology-driving trend and, indirectly, a CRT cost-driving trend because a higher yield is translated in lower CRT manufacturing costs and, consequently, lower cost per displayed megapixel.
- MIPS was chosen as a cost-per-megapixel-driving trend because, as the MIPS rating of the graphics microprocessor improves—incorporating an increase in its speed and a more efficient instruction set (see Equation 3.20)—the price per MIPS decreases (see actual data in Section 3.2 and ICs supply model results in Section 4.3). Thus, if the refresh rate, the number of colors displayed, and the resolution of the CRT are assumed to remain constant, the cost of the graphics adapter and, consequently, the cost per megapixel decreases.

²⁷Several monitors on the market offer varying operational bandwidths, due to the variety of graphics adapters available to drive the CRT.

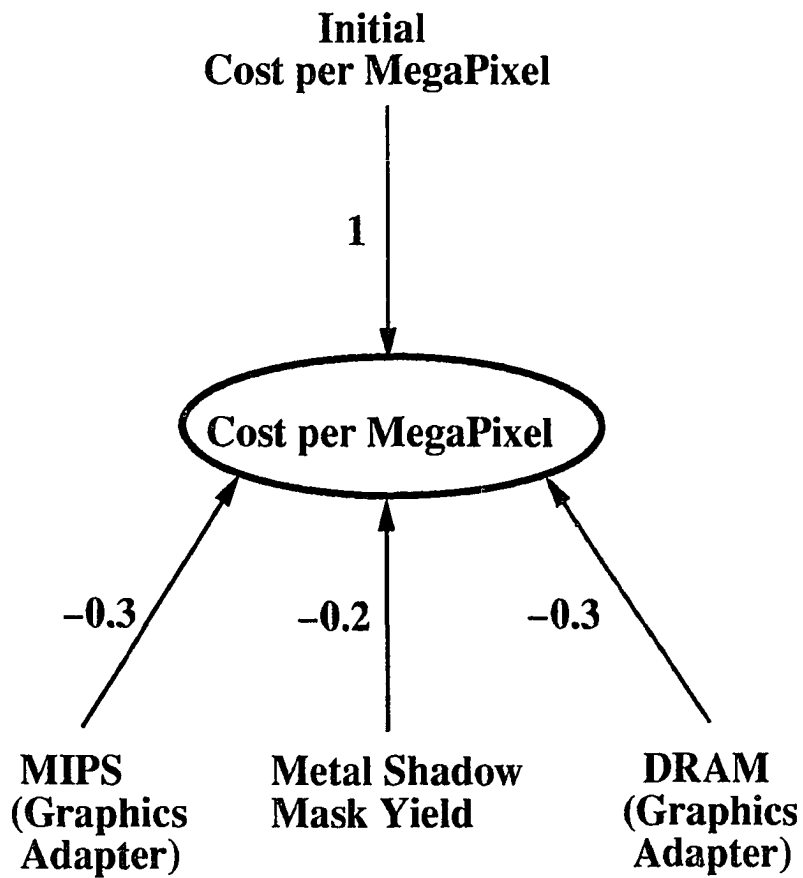


Figure 3.16: Relational diagram of the cost/megapixel of a color CRT display.

- DRAM capacity was chosen as a cost-per-megapixel-driving trend because, as the DRAM capacity of the graphics adapter increases, the DRAM price per megabyte decreases (see actual data in Section 3.2 and ICs supply model results in Section 4.3). Thus, if the refresh rate, the number of colors displayed, and the resolution of the CRT are assumed to remain constant, the cost of the graphics adapter and, consequently, the cost per megapixel decreases.

The cost per megapixel model can be expressed mathematically as follows:

$$CMP_t = CMP_0 * \left(\frac{MY_t}{MY_0}\right)^{-0.2} * \left(\frac{MIPS_t}{MIPS_0}\right)^{-0.3} * \left(\frac{MEMORY_t}{MEMORY_0}\right)^{-0.3} * 10^{-0.004*(t-t_0)} \text{ (\$/MP)} \quad (3.80)$$

where CMP_0 is equal to the value of Equation 3.69 evaluated at t_0 , MY represents the metal shadow mask production yield, $MIPS$ represents the number of million instruction per second that the graphics microprocessor can perform, $MEMORY$ represents the DRAM capacity installed on the adapter board, and t represents the year in which the cost per megapixel is calculated. Since most of the design yields and the effects of the apertures enhancements on the cost are kept as classified information [16], an exponentially decreasing factor, $10^{-0.004*(t-t_0)}$, is used in Equation 3.80 to capture these cost diminishing effects over time. The values of the exponents, (a_n s), in Equation 3.80 were obtained by tuning the model to yield results to match the actual cost per megapixel data provided in Subsection 3.4.2. All of the $|a_n|$ s are ≤ 1 , and each reflects an estimate of its relative importance to the behavior of the cost per megapixel over time. For example, a 10% increase in the metal shadow mask production yield leads to an

approximate decrease of 2% in the cost per megapixel (see Equation 3.9), and similarly for the other factors of Equation 3.80.

3.5 UNIX Operating System Supply Model

Software has always played a key role in the user's perception of a computer system. With more and more UNIX operated workstations sold and with more UNIX supported applications developed, UNIX has become the operating system of choice for many large-scale network distributed and multiprocessed applications [30]. The latest version of the UNIX operating system has more than 1 million lines of code [77]. With widespread support from the Open Software Foundation (OSF) and the engineering and business communities [30, 94], UNIX²⁸ has become the major operating system software in the industry.

This section is organized as follows: Subsection 3.5.1 presents the UNIX development-from-scratch and porting time periods and costs for both 1980 and 1991, Subsection 3.5.2 presents the assumptions and the terminology used in the development of the UNIX supply model, and Subsections 3.5.3 through 3.5.7 present the mathematical formulation of the UNIX operating system supply model, including equations that capture the time periods and costs of both the development-from-scratch and the porting of UNIX over the 1980-1991 study period.

²⁸The main attributes of UNIX that have enabled the computer workstation to achieve its current marketability are its optimal hardware resources utilization, single user multitasking, networking, and distributed computing support [77, 85].

3.5.1 UNIX Development-from-Scratch and Porting Trends

The times required for UNIX development-from-scratch and porting are company-dependent [72]. Developing software involves creativity and experience, two intangible and immeasurable attributes. To facilitate model development, some experienced UNIX developers at Uniforum [72] were consulted about development-from-scratch and porting issues related to the UNIX operating system. Based upon these discussions a model was formulated. The next two paragraphs report the UNIX development-from-scratch and porting time periods and costs data for 1980 and 1991. Unfortunately, the data for the years between 1980 and 1991 were not available.

1980 Development-from-Scratch and Porting: Time and Cost Data.

Since UNIX was not as popular in the early 1980s as it is today, the knowledge and experience of UNIX software engineers were not as sophisticated as is needed today. From the UNIX development specialists we learned that the UNIX development-from-scratch time period was nearly 15 man-years in 1980 and that the porting time was approximately 1.5 man-years [72]. The demand for UNIX software engineers far exceeded their availability in the late 1970s, and they charged a higher coding²⁹ cost per hour. The same UNIX development specialists estimated that the average coding cost per hour was close to \$70 in 1980 [72]. Based on the 1980 development-from-scratch and porting time periods and the average coding cost per hour, the development-from-scratch and

²⁹Coding is the job performed by the software engineer while developing the operating system's code; most of the UNIX code is written in the C programming language.

porting costs³⁰ can be computed as follows:

$$DFSC_{1980} = 15 * (70 * 8 * 250) = 2,100,000 \text{ (\$)} \quad (3.81)$$

$$PC_{1980} = 1.5 * (70 * 8 * 250) = 210,000 \text{ (\$)} \quad (3.82)$$

where $DFSC_{1980}$ is the 1980 development-from-scratch cost of UNIX, PC_{1980} is the 1980 porting cost of UNIX, 8 is the assumed number of working hours per day, and 250 is the assumed number of working days per man-year.

1991 Development-from-Scratch Time and Cost Data. Several workstation manufacturers like DEC, HP, IBM, and Sun Microsystems, Inc., have developed their own proprietary UNIX versions and packaged them in their computer systems. If the capabilities of UNIX were developed from scratch in 1991, it has been estimated that it would take only about 10 man-years [72]. When one considers also that in 1991 a software developer commanded an average salary of \$60,000 per year, and that the yearly job benefits and overhead totaled approximately \$40,000, the cost of developing the UNIX operating system from scratch would drop to nearly \$1 million [72]. Of this cost, the storage tape on which the UNIX software is kept costs no more than \$10 [72].

1991 Porting Time and Cost Data. UNIX porting costs are not as extensive as the development-from-scratch costs; nevertheless, it takes about 1 man-year to port UNIX to a different hardware platform [72] and costs close to \$100,000.

³⁰The development-from-scratch and porting costs reported are estimates, and the actual costs depend on the development company's overhead and its policy regarding the number of software prototypes, or beta-versions, that it tests before final release.

It is apparent from the previous three paragraphs that the UNIX development-from-scratch and porting time period and cost trends are decreasing. These trends reflect the much increased popularity of UNIX since the market introduction of the computer workstation and indicate how widespread porting and building new applications on top of the UNIX operating system is today.

3.5.2 Model Assumptions and Terminology

The operating system supply model incorporates equations capturing the time periods and costs required for the development-from-scratch and porting of the UNIX operating system. It costs no more than \$10 to package a copy of the UNIX software in a magnetic tape, so copying and packaging costs are negligible and will not be considered as an integral part of the supply model.

The main costs, after incurring the development-from-scratch or porting costs, involve software enhancements. Each company adds several enhancements to its own version to suit its machines and, depending on the demand for that company's machines, the price of the OS could vary from \$800 to \$2000 [86]. Before presenting the mathematical formulation of the model, several assumptions are listed. This is followed by a list defining the variables and the parameters used in the model.

Model Assumptions. In modeling the supply of developed-from-scratch or ported UNIX software, the following assumptions are used:

- The number of UNIX-based machines sold is increasing during the period of study [84, 85, 86].
- The number of UNIX development-from-scratch and porting consultants is increasing during the period of study [72].
- The average UNIX coding cost per hour is decreasing during the period of study [72].

Definitions and Terminology. In presenting the model's equations, several parameters and variables are introduced, and they are defined as follows:

0	the time index for the first year of the study period
t	the time index
SA	the software attributes index
WHA	the workstation hardware attributes index
$MIPS$	the number of million instructions per second
$MEMORY$	the capacity of a DRAM, in megabytes
DB	the bandwidth of the workstation display, in megaHertz
DR	the data rate of the workstation hard disk, in megabytes/second
UDS	the number of UNIX development specialists
$DUDS$	the demand for the UNIX development specialists
$ACCH$	the average coding cost per hour, in dollars/hour
$DFST$	the UNIX development-from-scratch time period, in years
PT	the UNIX porting time period, in years
$DFSC$	the UNIX development-from-scratch cost, in dollars
PC	the UNIX porting cost, in dollars

There are no standard software development tools and techniques for use by all software developers as of 1991. More than 50 books have been written on software development techniques [70] and the problems remain the same: how to make an *efficient* transition from the structural specifications of a software project to its actual completion, and how to assign a cost to a project from its structural specifications? The current software development paradigms do not allow these questions to be fully answered. What the operating system supply model attempts to capture is the overall effect of improved coding machines' hardware attributes on the time required to satisfy the structural specifications and on the cost associated with meeting these specifications.

The UNIX supply model is presented as follows: first, in Subsection 3.5.3 the UNIX development-from-scratch time period is presented; second, in Subsection 3.5.4 the UNIX porting time period is presented; third, in Subsection 3.5.5 the software attributes index is described and an assumed index behavioral trend introduced; fourth, in Subsection 3.5.6 the workstation³¹ hardware attributes index is described with all its factors and an index behavioral equation introduced; and fifth, in Subsection 3.5.7 the UNIX development-from-scratch and porting costs are computed.

3.5.3 UNIX Development-from-Scratch Time Period

Figure 3.17 illustrates, in a relational diagram, how software and hardware attributes affect the development-from-scratch time period of the UNIX

³¹ Assuming that a computer workstation is used to develop the software code.

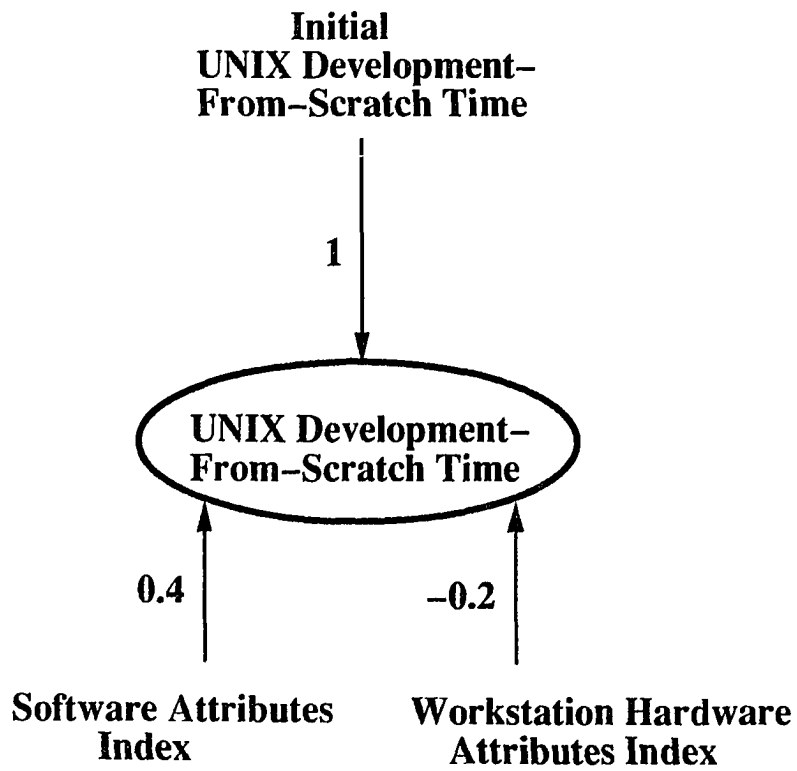


Figure 3.17: Relational diagram of the UNIX development-from-scratch time period.

operating system:

- Software attributes index was chosen as a UNIX development-from-scratch time-driving trend because, as the software attributes increase, the software specifications increase and, usually, the time required to meet all of them increases.
- Workstation hardware attributes index was chosen as a UNIX development-from-scratch time-driving trend because, as the hardware attributes of the coding machine³² become more advanced, the workstation's response time usually improves and, consequently, the software development task becomes more efficient.

The diagram in Figure 3.17 can be expressed mathematically as:

$$DFST_t = DFST_0 * \left(\frac{SA_t}{SA_0}\right)^{0.4} * \left(\frac{WHA_t}{WHA_0}\right)^{-0.2} \quad (year) \quad (3.83)$$

where $DFST_0$ represents the initial UNIX development-from-scratch time period. Equation 3.83 reflects how the UNIX development-from-scratch time period increases as the UNIX software attributes index (SA) increases, and decreases as the development workstation's hardware attributes index (WHA) increases. The values of the exponents, (a_n s), in Equation 3.83 were obtained by tuning the model to yield results to match the UNIX consultant estimated development-from-scratch trend provided in Subsection 3.5.1. All of the $|a_n|$ s are ≤ 1 , and each reflects an estimate of its relative importance to the behavior of the development-from-scratch period over time. For example, a 10% increase

³²Computer workstations are usually used in the software development process.

in the software attributes index leads to an approximate increase of 4% in the development-from-scratch time period (see Equation 3.9), and a 10% increase in the workstation hardware attributes index leads to an approximate decrease of 2% in the development-from-scratch time period.

3.5.4 UNIX Porting Time Period

The UNIX porting time period is affected by the same factors associated with the development-from-scratch of UNIX, and Figure 3.18 is the relational diagram showing how these factors might affect porting time. The diagram in Figure 3.18 can be expressed mathematically as:

$$PT_t = PT_0 * \left(\frac{SA_t}{SA_0}\right)^{0.4} * \left(\frac{WHA_t}{WHA_0}\right)^{-0.2} \quad (year) \quad (3.84)$$

where PT_0 represents the initial UNIX porting time period. The explanation of how each of Equation 3.84's factors and their exponents affect the porting time is similar to that provided for Equation 3.83 in the previous subsection.

3.5.5 Software Attributes Index

The software attributes of the UNIX operating system depend on the user's perception and, as a result, they are often difficult to measure. These attributes include the user friendliness of the software, coupled with its compatibility, networkability, portability, and efficiency. For modeling purposes, it is assumed that the index of all the software attributes presented earlier is increasing at a rate of 5% a year, or mathematically:

$$SA_t = SA_0 * 1.05^{t-1980}. \quad (3.85)$$

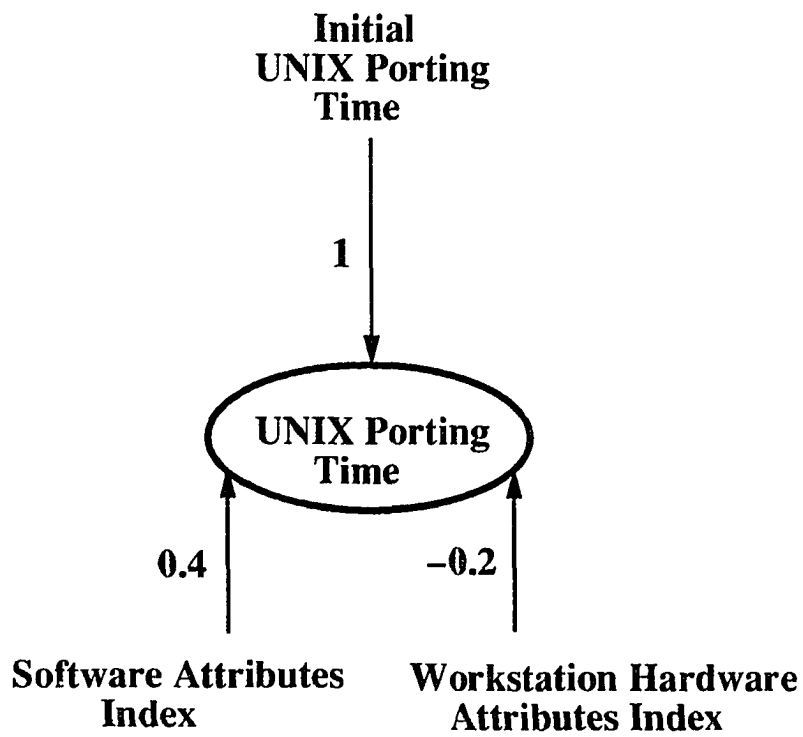


Figure 3.18: Relational diagram of the UNIX porting time period.

It does not matter what the initial value SA_0 is because only the relative change in the values of the software attributes index is relevant in Equations 3.83 and 3.84.

3.5.6 Workstation Hardware Attributes Index

Most computer hardware attributes are tangible entities, and their performance can be measured over time. Of these measurable attributes, the following were selected as potentially relevant to expressing a coding machine's capabilities:

- The MIPS rating of the workstation's integer unit was chosen because it reflects the processor's speed and architecture, and, partly, the workstation's throughput.
- The DRAMs capacity was chosen as a hardware attribute because denser DRAMs enable most of the software developer's code to remain in main memory during execution, thereby increasing the machine's response time and, in turn, reducing the development time.
- The CRT display bandwidth was chosen as a hardware attribute because it affects the size of the screen and, ultimately, its resolution. (It has been shown that larger screens improve the software development environment and make the coding process more efficient, and so reduce the development time.)
- The hard disk data rate was chosen as a hardware attribute because it reflects the hard disk data density and the rate at which the data are fed

back to the suspended program, while waiting on the data from the disk. (A fast data rate increases the disk's response time and, consequently, improves the efficiency of the code development process.)

All the above listed workstation hardware attributes are calculated in the ICs, magnetic hard disk, and CRT display supply models already reported in previous sections of this chapter. Use of these attributes integrates the models into one simple formulation. Figure 3.19 illustrates in a relational diagram the effects of the previously presented attributes on the overall index of the workstation's hardware attributes. The diagram in Figure 3.19 can be expressed mathematically as follows:

$$WHA_t = WHA_0 * \left(\frac{MIPS_t}{MIPS_0}\right)^{0.25} * \left(\frac{MEMORY_t}{MEMORY_0}\right)^{0.35} * \left(\frac{DB_t}{DB_0}\right)^{0.15} * \left(\frac{DR_t}{DR_0}\right)^{0.25} \quad (3.86)$$

The initial value WHA_0 of the workstation hardware attributes index is not relevant because only the relative change of the index's values matters in Equations 3.83 and 3.84. All of the a_n s add up to 1, and each reflects an estimate of its relative importance to the behavior of the hardware attributes index over time. For example, a 10% increase in the hard disk data rate leads to an approximate increase of 2.5% in the hardware attribute index (see Equation 3.9), and similarly for the rest of the factors of Equation 3.86.

3.5.7 UNIX Development-from-Scratch and Porting Costs

As shown earlier, the development-from-scratch and porting processes have associated with them long periods of coding time. The coding time spent has, in turn, an associated cost. This coding cost has decreased over the years

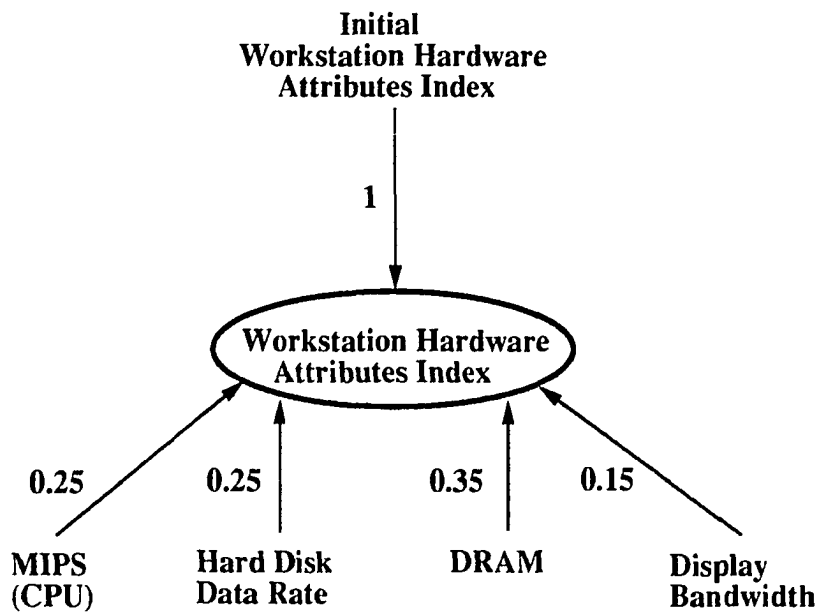


Figure 3.19: Relational diagram of the workstation hardware attributes.

because of the increased demand and supply of UNIX development specialists [72]. It is expressed mathematically as:

$$ACCH_t = ACCH_0 * \left(\frac{UDS_t}{UDS_0}\right)^{-1} * \left(\frac{DUDS_t}{DUDS_0}\right)^1 \quad (\$/hour) \quad (3.87)$$

where $ACCH_0$ is chosen to reflect the initial average coding cost per hour at t_0 . It is assumed that the number of UNIX development specialists and their market demand have been increasing yearly by 10% and 7%, respectively, since 1980, due to the proliferation of UNIX based workstations in the private and the public sectors. The UNIX development specialists trend is expressed mathematically as follows:

$$UDS_t = UDS_0 * 1.1^{t-1980} \quad (3.88)$$

and the market demand for UNIX development specialists is expressed as:

$$DUDS_t = DUDS_0 * 1.07^{t-1980}. \quad (3.89)$$

The values of UDS_0 and $DUDS_0$ in Equations 3.88 and 3.89 are immaterial because Equation 3.87 takes into consideration only the relative increase in the number of UNIX specialists and their demand and not their actual numerical values; t represents the year in which the number of UNIX development specialists and the demand for such specialists are calculated. From Equations 3.83, 3.84, and 3.87, the development-from-scratch and porting costs of the UNIX operating system can be computed as follows:

$$DFSC_t = ACCH_t * (8 * 250 * DFST_t) \quad (\$) \quad (3.90)$$

$$PC_t = ACCH_t * (8 * 250 * PT_t) \quad (\$) \quad (3.91)$$

where the $8 * 250$, or 2000, is equal to the number of working hours in one calendar man-year [72].

3.6 Workstation Assembly Model

Assembling a workstation has become a global effort. Most workstation assembly plants ship their raw materials and components from all over the world, and the assembled workstations are, in turn, shipped to many destinations worldwide. To be competitive, a workstation manufacturer must develop a mechanism to minimize its costs on a global scale, taking into consideration all the transportation, labor, and materials costs, the capacities of each plant's productive units, and the market demand each plant's output can satisfy. A useful conceptualization by which the manufacturing activity can be optimized is a process model, where the assembly plant is considered as a network with several processing stages, each stage with its specific input requirements and output levels. Each productive unit corresponds in the model to an assembly node with throughput bounds. Increasing the throughput bound requires allocation of investment capital by the manufacturer.

The workstation assembly network is illustrated in Figure 3.20, where components—CPUs, DRAMs, magnetic hard disks, CRT displays, and UNIX operating systems, etc.—are assembled into a final product. But each of the components is itself an intermediate product. These first intermediate products are the CPU boards, the DRAM boards, the magnetic hard disks with their interfaces to the computer systems, the CRT displays with their interfaces to the computer systems, and the UNIX operating systems with their installation kits ready for loading. At a second intermediate stage, a CPU board, a DRAM board, a magnetic hard disk with its interface, and an electric power supply are assembled into a workstation box. At a third intermediate stage, the operating system is loaded in the hard disk, and the keyboard, mouse, and display with

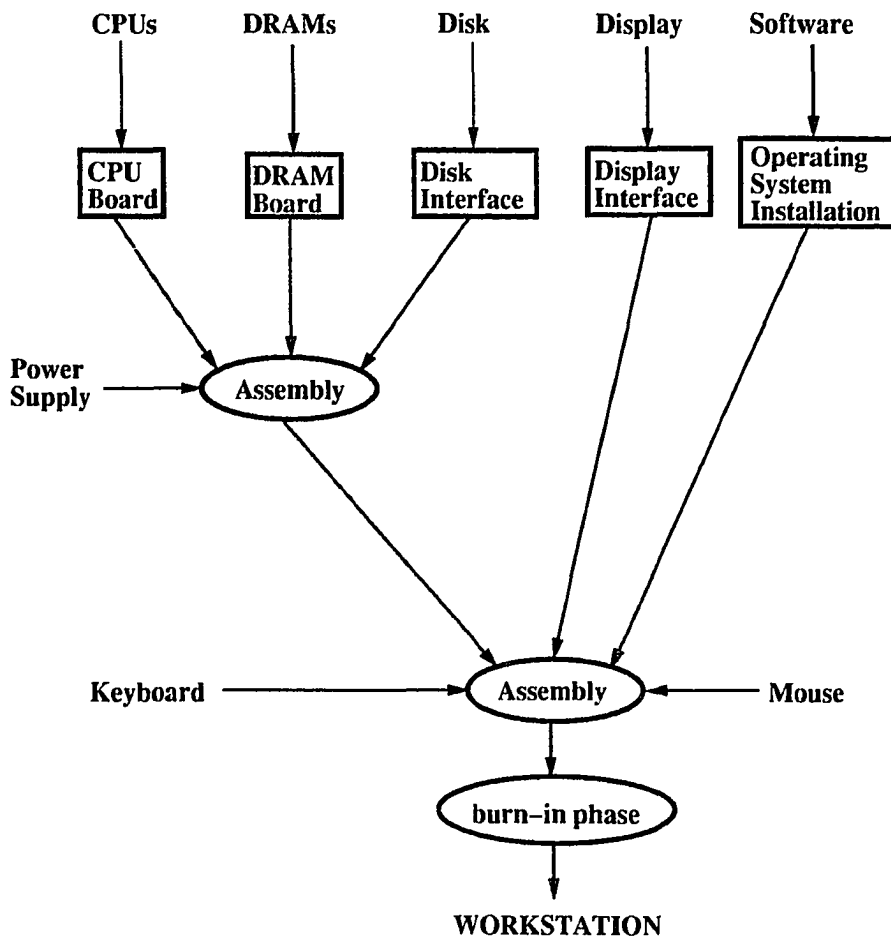


Figure 3.20: Illustration of a workstation assembly network.

its interface are connected to the box's corresponding external ports³³. The assembled workstation goes through a final burn-in stage before it is ready for shipping to the distribution centers. Each of the assembly points in the network has a capacity limit that can be expanded only by increasing the plant's space, machinery, raw materials inventory, and labor.

The developed linear process model provides an optimal allocation of inputs among the different activities to minimize the assembly costs over several intervals³⁴ of time. The model is a simple illustration of the workstation assembly process, and only one manufacturing plant is considered, with its market demand located in its vicinity (i.e., no workstations transportation costs are considered). Minimizing the costs is the model's objective function. The cost minimization is constrained by internal capacity limitations and the market requirement for finished workstations that must be satisfied. Though no import, export, plant expansion, and government quota restrictions are included, conceptually there is no difficulty in adding them. There is extensive literature on the use of process models to analyze problems of the type considered here. For more details, see Kendrick/Stoutjesdijk [46] or Kendrick/Meeraus/Alatorre [45].

The presentation of the workstation assembly model will proceed as follows: first, in Subsection 3.6.1 the workstation assembly steps are described; second, in Subsection 3.6.2 the assumptions and terminology used in the development of the workstation assembly model are presented; and third, in Subsection 3.6.3 the model is formulated.

³³A port is a physical connection to the machine's internal hardware and serves as a communication gate between the processing boards and the peripherals.

³⁴An interval is usually one year because the available data is provided annually.

3.6.1 Assembly Steps

The assembly process in Figure 3.20 can be described in five steps:

1. The components meeting the specifications of the configured machine are prepared at the machine's assembly site.
2. The CPU and the DRAM chips are mounted on the printed circuit boards—i.e., the processing board and the main memory (DRAM) board.
3. The processing board, the main memory board, the hard disk storage with its interface, and the electric power supply are grouped and tested for correct assembly.
4. If the already assembled hardware is functional, the operating system is loaded into the hard disk, and the keyboard, mouse, and display with its interface are connected to the machine.
5. A burn-in phase takes place where the functionality of the workstation is tested for a period of 10 to 24 hours.

3.6.2 Model Assumptions and Terminology

In formulating the linear process model the following assumptions were used:

- The arcs in the network are directed.
- The input amount of each component in the assembly process is constant during one time period.

- The price of each component is constant during one time period.
- Some of the hardware components are produced outside the United States. However, due to the lack of data on the geographical locations of the components manufacturers, it will be assumed that all the components needed for assembly will be available at the workstations assembly site. The cost of transporting these components to the assembly plant will be incorporated in their prices.
- Each CPU die incorporates an integer unit (IU), a floating point unit (FPU), a memory management unit (MMU), and a cache memory. The price of the CPU reflects the price of the set.
- Only one assembly plant is considered.
- The demand is assumed to be concentrated in the vicinity of the production plant; no workstations transportation costs are considered.
- No plant expansion costs are considered, and the market demand for the workstations produced from the plant remains constant at a value less than or equal (\leq) to the plant's production capacity.
- The monetary discount rate is assumed to be 5%, compounded yearly. The discount rate does account for the inflation of prices during the study period.

Definitions and Terminology In presenting the linear process model's equations, several parameters and variables are introduced, and they are defined as follows:

t	an index to the time periods T
c	an index to the commodities C, CR, CI & CF
p	an index to the set of processes P
m	an index to the set of productive units M
DR	the discount rate, in %
A_{cp}	the input-output coefficients matrix
B_{mp}	the plant capacity utilization matrix
K_{mt}	the capacities of the productive units matrix
DD	the demand distribution matrix, in %
$PRICES_{ct}$	the price matrix of the components and raw materials
z_{pt}	the process level variable
x_{ct}	the amount of final products shipped variable
u_{ct}	the amount of materials purchased variable
$RMATC_t$	the cost variable of the components and raw materials

3.6.3 Model Formulation

The **objective** function in the linear model is to minimize the present value of the total workstation components and raw materials costs, where future costs are discounted using the market rate:

$$\text{MIN} \sum_{t=1}^T \left(1 + \frac{DR}{100}\right)^{(1-t)} * \text{RMATC}_t. \quad (3.91)$$

This model deals with the assembly aspect of workstations only. For instance, no transportation costs are included in Equation 3.91 because we assumed that the demand is in the vicinity of the production plant. No expansion costs are included because we assumed that the demand is constant over the period of

study.

The workstations components and raw materials costs equations are equal to the products of the prices provided on Table 4.16 and the amount of components and raw materials required in the production process:

$$\mathbf{RMATC}_t = \sum_{c=1}^{CR} \mathbf{PRICES}_{ct} * \mathbf{u}_{ct} \quad \text{for } t = 1 \cdots T. \quad (3.93)$$

Since no component imports or product exports are considered, the raw materials balance constraints at the production plant are:

$$\sum_{p=1}^P \mathbf{A}_{cp} * \mathbf{z}_{pt} \geq -\mathbf{u}_{ct} \quad \text{for } c = 1 \cdots CR, t = 1 \cdots T. \quad (3.94)$$

The intermediate materials balance constraints are:

$$\sum_{p=1}^P \mathbf{A}_{cp} * \mathbf{z}_{pt} \geq 0 \quad \text{for } c = 1 \cdots CI, t = 1 \cdots T. \quad (3.95)$$

And the final materials balance constraints are:

$$\sum_{p=1}^P \mathbf{A}_{cp} * \mathbf{z}_{pt} \geq \mathbf{x}_{ct} \quad \text{for } c = 1 \cdots CF, t = 1 \cdots T. \quad (3.96)$$

The capacity constraints at the assembly plant require that the plant's output cannot exceed the capacity of its production units:

$$\sum_{p=1}^P \mathbf{B}_{mp} * \mathbf{z}_{pt} \leq \mathbf{K}_{mt} \quad \text{for } m = 1 \cdots M, t = 1 \cdots T. \quad (3.97)$$

The demand constraints guarantee that the number of workstations produced from the plant satisfy the market demand:

$$\mathbf{x}_{ct} \geq 10,000 * \frac{\mathbf{DD}}{100} \quad \text{for } c = 1 \cdots CF, t = 1 \cdots T \quad (3.98)$$

and since we are assuming the demand is generated by a single market, \mathbf{DD} is equal to 100%.

There are several non-negativity constraints, including:

$$\mathbf{z}_{pt}, \mathbf{x}_{ct}, \mathbf{u}_{ct} \geq 0 \quad \text{for } c = 1 \dots C, p = 1 \dots P, t = 1 \dots T. \quad (3.99)$$

The linear process model was coded in GAMS [11], and the BDMLP³⁵ solver on an IBM 3081KX VM/XA mainframe was used to solve it.

³⁵BDMLP is a linear programming solver.

Chapter 4

Model Behavior and Sensitivity Results

The workstation component supply models presented in Chapter 3 are discrete event simulation models. Each model was designed and tuned to capture certain component trends over time. This chapter illustrates the dynamic behavior of the component supply models. This is not a validation of the models. A true validation, in a scientific sense, is only possible by using the models to test hypotheses. In other words, a model prediction needs to be made (a hypothesis formed), and if the prediction turns out to be true, the model passes the validation test.

More important than model validity in the strict scientific sense is how useful the model is at helping us to understand and to gain insight into past and possible future trends. To illustrate the models' utility in this regard, we present a base case scenario of future trends for each of the workstation components and their assembly into completed workstations. This is followed by sensitivity analyses which make clear the strength of the models' linkages. As will be seen, this set of analyses indicates that the supply models developed in Chapter 3 can yield practical insights into the relationships among a workstation's components, capabilities, and costs.

A careful examination of how the model results over a past period of time compare with actual market data over the same historical period is

presented. This discussion, however, simply shows how well the supply models of Chapter 3 were tuned to capture the historical behavior of the data. A stronger form of comparison is simply not possible with the current models and the actual available data. The closer the models' results are to the actual market data, the better tuned the models are, unless the actual market data represents company-dependent decisions that the models cannot capture.

The supply models were tuned to capture unit market price trends of the various workstation components and the behavior of certain component capabilities over time. The unit price results of the component supply models include the CPU and DRAM prices, the price per megabyte of magnetic hard disk, the price per megapixel of color CRT display, and the UNIX porting and development-from-scratch prices. The component capabilities include the operational speed of the microprocessor, the capacity of the main memory, the capacity of the magnetic hard disk, the size and resolution of the color CRT display, and the functionality of the operating system.

Most of the market data collected reflect the attributes of the components during the 1980s; consequently, the components supply models¹ were simulated and their results compared to actual data over the 1980-1991 period. The simulation code of the component supply models was written in C, and an IBM RISC System/6000 POWERserver/530 was used to obtain the simulation results.

¹The supply models reflect only the attributes of the components already on the market, not those available in the manufacturers laboratories. Most of the future ICs and computer systems products are provided in the IEEE International Solid-State Circuits Conference-ISSCC Digest of Technical Papers and the IEEE COMPCON, respectively.

This chapter is organized as follows:

- Section 4.1 presents the cost markup percentage used in the supply models to obtain the component prices, and illustrates the difference between the single unit and the bulk price of a component.
- Section 4.2 lists the component supply models' input parameters and the rates of change of each of the components' physical characteristics, capabilities, and price trends, if available.
- Sections 4.3 through 4.6 present and compare the results of the supply models and the actual market data of the IC components, the magnetic hard disk, the color CRT display, and the UNIX operating system.
- Section 4.7 presents the workstation assembly model inputs and results.
- Section 4.8 performs sensitivity analyses of the workstation components and assembled workstations to the ICs' feature size and the number of silicon wafer defects per unit area.

4.1 Component Cost, Single Unit Price, and Bulk Price

Since the supply models provide single unit component costs, not prices, as their output, all of the cost results were marked up by 200% to compare them with the available market price data [33]:

$$\textit{Single Unit Price}_{model} = \textit{Single Unit Cost}_{model} * (1 + 200\%) \text{ (\$)}. \quad (4.1)$$

The 200% cost markup percentage includes the gross margin and the average discount percentages that the manufacturers add to their products' costs².

Most manufacturers sell their products in bulk, in quantities larger than or equal to 10,000 units. The bulk unit price is usually equal to half the single unit list price [16]:

$$\text{Bulk Unit Price} = \frac{\text{Single Unit Price}}{2} \quad (\$). \quad (4.2)$$

Since the actual data collected are single unit list prices, though, the results of the models are presented in the following sections as single unit list prices.

4.2 Component Supply Model Inputs

Each of the component supply models has a set of input parameters, the values of which can be changed interactively while executing the simulation program. As presented in Chapter 3, certain CPU and DRAM attributes of the ICs model, like the MIPS of CISC CPUs and the capacity of DRAMs, were incorporated as factors influencing the attributes of other models; the inputs of the ICs supply model are, then, a part of the input parameters of all the supply models.

This section presents the input parameters of the component supply models in Subsection 4.2.1, and lists the rates of change of the components physical characteristics trends in Subsection 4.2.2 and the rates of change of the components capabilities and price trends in Subsection 4.2.3.

²Some manufacturers mark up their components' costs by as much as 300% [33].

4.2.1 Model Input Parameters

The list below presents the values of the input parameters of the component supply models for the first year of the study period, or for the first year of the component's market introduction. For instance, the RISC architecture was widely marketed in 1987 [84], and the 19-inch computer color CRT display was introduced in 1985 [16]; their input parameters are given for the years 1987 and 1985, respectively. Even though the study period spans from 1980 until 1991, the RISC CPUs and the color CRT display models will be simulated only from 1987 to 1991, and from 1985 to 1991, respectively.

Simulation Period

Starting Year = 1980

Study Period = 1980-1991

ICs Supply Model Inputs

1980 - Wafer Cost = \$350

1980 - Wafer Yield = 90%

1980 - Number of Silicon Wafer Defects per unit area = 2.5 defects/cm²

1980 - Number of Test-Dies per Wafer = 2 dies

1980 - Testing Cost per Hour = \$210/hour

1980 - Average Die Test Time = 15 seconds

1980 - CISC CPUs Speed per cm² = 25 megaHertz/cm²

1987 - RISC CPUs Speed per cm² = 28 megaHertz/cm²

1980 - DRAM Density = 0.025 megabyte/cm²

Magnetic Hard Disk Supply Model Inputs

1980 - Head-Actuator Setup Width = 1.25 cm

Color CRT Display Supply Model Inputs

1985 - Screen Refresh Rate = 70 Hertz

1985 - Number of Metal Shadow Mask Defects per Inch = 1 defect/inch

UNIX Operating System Supply Model Inputs

1980 - Average Coding Cost per Hour = \$70/hour

1980 - UNIX Porting Time Period = 1.5 man-years

1980 - UNIX Development-from-Scratch Time Period = 15 man-years

4.2.2 Components' Physical Characteristics Trends

The list below presents the rates of change of the components' physical characteristics trends during the study period, if available.

ICs Supply Model

CISC CPUs die areas: 7.53% per year exponential growth

DRAMs die areas: 7.53% per year exponential growth

Feature size: 5.5% per year exponential decline

Number of masking levels: 8% per year exponential growth

Silicon Wafer Diameter: 2.52% per year exponential growth

Magnetic Hard Disk Supply Model

Magnetic bit cell length: 4.9% per year exponential decline

Magnetic disk track pitch: 6.6% per year exponential decline

Magnetic medium thickness: 2.95% per year exponential decline

Magnetic head gap spacing: 4% per year exponential decline

Head-medium spacing: 5.5% per year exponential decline

Head-actuator setup width: 4% per year exponential decline

Color CRT Display Supply Model

Metal shadow mask hole pitch: 4% per year exponential decline

4.2.3 Components' Capabilities and Price Trends

The list below presents the rates of change of the components' capabilities and price trends during the period of study, if available.

ICs Supply Model

CISC CPUs IPC: 11% per year exponential growth

RISC CPUs IPC: 6.3% per year exponential growth

Magnetic Hard Disk Supply Model

Price per megabyte: 8.57% per year exponential decline

Areal density: 13.8% per year exponential growth

Color CRT Display Supply Model

Price per megapixel: 8.5% per year exponential decline

4.3 ICs: Model Results and Actual Market Data

The following section presents the ICs supply model results and compares them to the actual ICs market data presented in Section 3.2. In Subsection 4.3.1 the model results and the estimated market data of the ICs die yields are presented and compared; in Subsections 4.3.2, 4.3.3, and 4.3.4 CISC CPUs, RISC CPUs, and DRAMs model results and actual market data are presented and compared.

4.3.1 ICs Die Yields

Column 1 of Table 4.1 lists some common die areas of manufactured ICs, and Columns 2 and 3 present the corresponding 1989 model results on die yields and actual market data³ for the various die sizes. The values in Column 2 and Column 3 differ by less than 10%.

Most of the die yields data are classified as confidential information by the manufacturers because they reflect their positions on the learning curves and their percentages of profit. If made public, the die yields information can be used by competitors to create strategies to gain market share from the manufacturer whose die yields information was disclosed.

4.3.2 ICs: CISC CPUs Model Results and Actual Market Data

The following paragraphs compare the model results and the actual market data of CISC microprocessors. MIPS ratings, prices, price/MIPS, and speed per cm² of CISC CPUs are provided.

CISC CPUs MIPS and Prices. Table 4.2 presents a comparison of selected model results with actual data of Intel CISC microprocessors. Only Intel's data was considered because Motorola's CISC CPUs price data was not available. For each Intel CPU die area, the MIPS rating and the price are computed, and compared to the data already presented on Tables 3.2 and 3.3 of Chapter 3. The

³The market data were estimated in Chapter 2 of Hennessy/Patterson [33], and were presented in Section 3.2 of this dissertation.

ICs : Die Yields		
1	2	3
Die Area (cm ²)	Die Yield (Model Results) (%)	Die Yield (Actual Data) (%)
0.0625	77.3	78
0.2601	49.8	46
0.5776	27.3	22
1.0404	13.4	10
1.6129	6.6	5
2.3104	3.3	3
3.1684	1.6	2
4.1209	0.9	1

Table 4.1: Model results and actual market data of die yields for certain die areas in 1989. Source: [33].

ICs: CISC CPUs						
1	2	3	4	5	6	7
Year	CPU Model #	Area (cm ²)	MIPS (Model Results)	MIPS (Actual Data)	Price (Model Results) (\$)	Price (Actual Data) (\$)
1982	i80286	0.60	0.82	2	67	360
1985	i80386DX	1.13	3.73	3	330	299
1989	i80486DX	1.65	17.95	11.40	552	700
1991	i8086	0.28	0.38	0.33	35	1.50
1991	i80286	0.60	1.98	2	61	7
1991	i80386DX	1.13	9.11	8.30	321	166
1991	i80486DX	1.65	32.60	35	558	588

Table 4.2: Model results and actual market data on MIPS ratings and prices of Intel CISC CPUs. Sources: [41, 42, 48].

MIPS⁴ model results and the actual data in Columns 4 and 5 of Table 4.2 differ, on average, by less than 25%.

Some of the CPU model price results and actual price data on Table 4.2 differ significantly for certain CPUs. The actual market price of the i80286 in 1982 differs from the model result by almost \$300. This is probably because, at the time, the Intel processors had gained wide acceptance for use in the IBM PC. What enhanced the i80286's acceptance in 1982 was IBM's decision to make public its bus and Intel-based processor board architectures to compete with and perhaps capture Apple Computer's market share of PCs. As a result, the demand for the i80286s increased and Intel took a monopolistic stance, charging a sizable markup—larger than 200%—since it was the sole producer of that chip. By 1985, when the first i80386s were introduced, the model results and the actual data compare much more closely.

The difference between the 1991 actual market price data of the i8086/286/386DX ICs and the model price results can be attributed to the fact that the old technologies were nearing obsolescence. As new technologies are introduced, the demand for old ones diminishes, forcing their prices down. After a while, the i8086/286/386DX market prices were even depressed below a fully marked up production cost. Moreover, in 1991 Intel wanted to sell its stock of old CPUs to reduce their high inventory costs, which stimulated more price cuts.

⁴The MIPS data and results reflect the architectural and production improvements to the CPU itself only, and not the entire chip set where it is used. Hence, the MIPS rating indicates solely the relative performance difference from other ICs in its class.

CISC CPUs Price/MIPS. Table 4.3 compares the price per MIPS model results with the actual price per MIPS data of CISC CPUs. The entries on Table 4.3 were obtained by dividing the price of each CPU on Table 4.2 by its MIPS rating. The model results and actual market data reveal a trend of decrease in the price per MIPS due to architectural design enhancements of CISC CPUs and higher manufacturing yields. The average error is less than 28%.

CISC CPUs Speed/cm². Table 4.4 lists the model results and the actual market data on the operational speeds per unit die area of CISC CPUs (unavailable data is denoted by n.a.). The average difference between the actual data and the model results is less than 15%. The model results range from 25 to 40.4 megaHertz per cm² in the 1980-1991 period, while the market data are scattered in the 24-40.8 megaHertz per cm² range during the same period. Some of the values in Columns 2 and 3 of Table 4.4 differ significantly, and this is due to the different chip designs and layouts used by the manufacturers during the period of study. As mentioned in Chapter 2, a simpler CPU design always improves the speed because the chip layout becomes easier and the inner-chip communication time decreases due to the enhanced and optimized layout.

4.3.3 ICs: RISC CPUs Model Results and Actual Market Data

The following paragraphs compare the model results and the actual market data of RISC microprocessors. Where available, MIPS ratings, prices, and speed per cm² of RISC CPUs are provided.

ICs: CISC CPUs		
1	2	3
Year	Price/MIPS (Model Results) (\$)	Price/MIPS (Actual Data) (\$)
1982	81.9	180
1985	88.5	99.6
1989	30.6	61.4
1991	17.1	16.8

Table 4.3: Model results and actual market data on the price/MIPS of Intel CISC CPUs. Sources: [41, 42].

ICs: CISC CPUs		
1	2	3
Year	Speed/cm ² (Model Results)	Speed/cm ² (Actual Data)
1980	25.00	n.a.
1981	25.97	n.a.
1982	27.29	30.3
1983	28.35	24.4
1984	29.76	n.a.
1985	30.91	33.8
1986	32.42	n.a.
1987	33.97	40.8
1988	35.28	29.2
1989	36.94	n.a.
1990	38.66	30.3
1991	40.43	40

Table 4.4: Model results and actual market data on the speed/cm² rating of CISC CPUs. Sources: Intel data [41, 42, 48], Motorola data [61, 71, 84].

RISC CPUs MIPS and Prices. As mentioned earlier, the ICs model simulation period for the RISC⁵ CPUs starts in 1987, instead of 1980, and ends in 1991. Hewlett-Packard's Precision Architecture (HP-PA) RISC CPUs data is compared with the model results because their die areas were the only actual data available. The MIPS ratings provided on Table 4.5 for the HP-PA CPUs are those of the integer units of the processing chip sets used in the HP workstations or servers [22, 37]. The actual MIPS ratings data (Column 5) are increasing at a faster rate than the model results (Column 4). This may be due to our lack of knowledge of the actual HP's manufacturing processes. One explanation for such an improvement in performance is related to the transfer of HP's mainframe technologies to the workstation⁶. No HP RISC CPUs price data was available because most of the HP chips are proprietary.

RISC CPUs Speed/cm². Table 4.6 lists the model results and the HP-PA market data on the operational speeds per unit die area of RISC CPUs. Again, the model results are increasing at a slower pace than those of the market data for the same reasons as given in the previous paragraph. The model results show an increase from 28 to 33.96 MHz/cm² in the 4-year period, while the HP-PA CPU operational speed per unit die area actually more than doubled in a period of 2 years from 15.3 to 33.67 MHz/cm².

⁵The RISC CPUs die areas and operational speeds are larger than the CISC's because the RISC architecture is simpler than CISC's, and the RISC CPU layout is much easier to manufacture and has a higher yield than a similar die area CISC CPU yield.

⁶Since the workstation market has been eating away at the mainframe's market share for the last ten years, HP might have deemed it necessary to capture a large share of the fastest growing sector of desktop and networkable computers—the workstation sector—while foregoing larger short term profits that could have been made from selling mainframes.

ICs: RISC CPUs						
1	2	3	4	5	6	7
Year	System	Die Area (cm ²)	MIPS (Model Results)	MIPS (Actual Data)	Price (Model Results) (\$)	Price (Actual Data) (\$)
1987	HP9000/825	1.33	23	9	404.91	n.a.
1988	HP9000/835	1.58	32	14	507.36	n.a.
1989	HP9000/845	1.96	49	22	787.80	n.a.
1991	HP9000/730	1.96	71	76	804.81	n.a.

Table 4.5: Model results and actual market data on MIPS ratings and prices of HP RISC CPUs. Sources: [7, 22, 37, 51, 53, 89, 103].

ICs: RISC CPUs		
1	2	3
Year	Speed/cm ² (Model Results)	Speed/cm ² (Actual Data)
1987	28.00	n.a.
1988	29.64	n.a.
1989	31.03	15.3
1990	32.47	n.a.
1991	33.96	33.67

Table 4.6: Model results and actual market data on the speed/cm² rating of RISC CPUs. Sources: HP data [7, 22, 37, 51, 53, 89, 103], Sun data [84, 86].

4.3.4 ICs: DRAMs Model Results and Actual Market Data

The following paragraphs compare the model results and the actual Motorola market data of DRAMs. Capacities, capacity per cm^2 , prices, and price per megabyte of DRAMs are provided.

DRAMs Capacities and Prices. Table 4.7 presents the model results and the actual Motorola market data on DRAM capacities and prices. For each DRAM die area, the chip's memory capacity and price are computed, and the results are listed in Columns 4 and 6 of the table. The model results on DRAM capacities and the actual Motorola market data on Table 4.7 are almost identical. However, the 1984 model price results and actual data differ significantly because the old technologies were nearing obsolescence. As new technologies are introduced the demand for old ones diminishes, forcing their prices down. After a while, their market prices are even depressed below a fully marked up production cost.

The price differences in 1986 and 1988 show that either Motorola used rather large markups during their initial pricing of the 1 and 4 megabit DRAMs or they could not achieve high enough die yields to price their DRAMs lower. It has been suggested that the Motorola DRAM prices presented on Table 4.7 left a huge market window for the Japanese ICs manufacturers to infiltrate the DRAM market with prices far below Motorola's—a phenomenon known at the time as “market memory dumping and flooding.” The Motorola 1 and 4 megabit DRAM prices were cut by more than 50% six months after their market introduction in 1986 and 1988 [62], so perhaps there is something to the suggestion.

ICs: DRAMs						
1	2	3	4	5	6	7
Year	DRAM (bits)	Die Area (cm ²)	Capacity (Model Results) (MB)	Capacity (Actual Data) (MB)	Price (Model Results) (\$)	Price (Actual Data) (\$)
1984	16K	0.20	0.005	0.002	29	1.09
1984	64K	0.24	0.008	0.008	31	3.4
1984	256K	0.40	0.043	0.032	42	17.9
1986	1M	0.60	0.13	0.125	62	100
1988	4M	0.80	0.34	0.5	88	264
1991	16M	1.34	1.91	2	396	329

Table 4.7: Model results and actual Motorola market data on DRAM capacities and prices. Sources: [48, 62].

DRAM Price/MB. Table 4.8 presents the model results and the actual Motorola market data on the DRAM price per megabyte. The model results and the actual market data presented in Columns 2 and 3 further illustrate possible Motorola overpricing or low production yields. It was not until 1991 that the actual price per megabyte of DRAM was reduced enough to be consistent with the model results.

DRAM Capacity/cm². Table 4.9 presents the model results and the actual Motorola market data on the number of DRAM megabytes per unit die area. The model results and the market data are very similar. The trend reflects a consistent increase in the DRAM density, partly traceable to the trend of decrease in feature size. As the DRAM density increases and the price per megabyte decreases, semiconductor DRAMs might replace the magnetic hard disk as the main computer storage technology, as soon as nonvolatile ones are developed. DRAMs are faster and more reliable. (In Subsection 4.8.1 of this dissertation, the ICs and the magnetic storage supply models are used to illustrate a case wherein the price per megabyte of DRAMs becomes cheaper than that of magnetic storage by the turn of the century.)

4.4 Magnetic Storage: Model Results and Actual Market Data

Model results and actual market data on magnetic hard disk price per megabyte and areal densities are compared in Subsections 4.4.1 and 4.4.2. Subsection 4.4.3 explains why magnetic hard disk volumetric density and data rate results from the model are not compared to actual market data.

ICs: DRAMs		
1	2	3
Year	Price/MB (Model Results) (\$)	Price/MB (Actual Data) (\$)
1984	1060.5	559.3
1986	476	800
1988	260	528
1991	207	164.5

Table 4.8: Model results and actual Motorola market data on the price/megabyte of DRAMs. Sources: [48, 62].

ICs: DRAMs		
1	2	3
Year	MB/cm ² (Model Results)	MB/cm ² (Actual Data)
1980	0.02	n.a.
1981	0.03	0.034
1982	0.05	n.a.
1983	0.07	n.a.
1984	0.11	0.08
1985	0.14	n.a.
1986	0.22	0.21
1987	0.33	n.a.
1988	0.43	0.625
1989	0.65	n.a.
1990	0.96	n.a.
1991	1.43	1.49

Table 4.9: Model results and actual Motorola market data on the number of DRAM megabytes/cm². Sources: [48, 62].

4.4.1 Magnetic Hard Disk Price/MB

For more than thirty years, the price per megabyte of magnetic hard disk has been decreasing. Several factors have contributed to the decrease, including the increase in the manufacturing yields of magnetic hard disks with smaller track pitches and bit cell lengths and the decrease in the prices of fast CISC CPUs and high capacity DRAMs.

Table 4.10 presents in Columns 2 and 3 the model results and the actual data on the price per megabyte of magnetic hard disk. The model results and the actual market data differ by less than 10% for each corresponding year, and an examination of the trends reveals that the values dropped by almost 90% over a period of 11 years.

4.4.2 Magnetic Hard Disk Areal Density

Table 4.11 presents in Columns 2 and 3 the model results and the actual market data on the magnetic hard disk areal densities. Most of the model results in Table 4.11's second column are slightly larger than the market data in the third column, but by less than a 12% margin. Note that the actual areal density of magnetic storage in 1991 is more than 34 times larger than the actual areal density in 1980, while the actual price per megabyte of magnetic storage in 1991 is only about 1/9th that of 1980. This suggests that in the 1980s magnetic hard disk manufacturers were pressed to achieve higher magnetic storage densities to compete with optical storage. Furthermore, they could afford not to keep their prices apace with areal density improvements because magnetic storage response

Magnetic Hard Disks		
1	2	3
Year	Price/MB (Model Results) (\$)	Price/MB (Actual Data) (\$)
1980	56.23	56.23
1981	46.21	46.16
1982	37.52	37.90
1983	30.83	31.11
1984	25.07	25.54
1985	20.60	20.97
1986	16.76	17.21
1987	13.66	14.13
1988	11.22	11.60
1989	9.15	9.52
1990	7.46	7.82
1991	6.09	6.42

Table 4.10: Model results and actual market data on the price/megabyte of magnetic hard disks. Source: [88].

Magnetic Hard Disks		
1	2	3
Year	Areal Density (Model Results) (MB/cm ²)	Areal Density (Actual Data) (MB/cm ²)
1980	0.16	0.09
1981	0.21	0.13
1982	0.27	0.18
1983	0.35	0.25
1984	0.46	0.34
1985	0.60	0.46
1986	0.78	0.64
1987	1.01	0.88
1988	1.32	1.20
1989	1.72	1.65
1990	2.24	2.27
1991	2.92	3.12

Table 4.11: Model results and actual market data on the areal density of magnetic hard disks. Source: [57].

time is much shorter than optical, and its price per megabyte is much less⁷.

4.4.3 Notes on Volumetric Density and Data Rate

Volumetric density depends on hard disk design factors like the number of coated disk surfaces and the head-medium setup height, which affects the number of disks per storage device height. Since these factors are manufacturer-dependent, no collective volumetric density market trends are available, and no model results will be presented for comparison.

The hard disk *data rate* is itself dependent on the hard disk rotational speed and the number of disk surfaces per unit storage component height. Here, too, no general market trend for the data rate was available. Consequently, no comparative model results are provided.

4.5 Color CRT Display: Model Results and Actual Market Data

The 19-inch color CRT display was introduced in 1985 and became the principal workstation and CAD display [16]. Two color CRT mask technologies have dominated the market, the Sony Trinitron and the small holes perforated metal shadow mask technologies. In the display model, only the technology trends for metal shadow masks perforated with small holes were incorporated.

⁷The suggestion does not exclude the large R&D costs the magnetic storage manufacturers have to pay to achieve the high areal density; however, these costs are not so high to keep the magnetic storage price per megabyte rate of decrease at approximately 1/3rd the rate of increase of areal density.

Model results and actual market data on the color CRT price per megapixel and number of pixels per inch are presented in Subsections 4.5.1 and 4.5.2, respectively.

4.5.1 Color CRT Price/Megapixel

Table 4.12 presents in Columns 2 and 3 the model results and the actual market data on the price per megapixel of a 19-inch color CRT display. Most of the values in the two columns differ by less than 10% for each corresponding year. It is interesting to note that a 19-inch color CRT display in 1985 was priced as much as an average computer workstation in 1991. Very few single users could afford such displays. Nevertheless, with improved CRT and metal shadow mask manufacturing yields, and decreased costs of CISC CPUs and DRAMs, the CRT prices dropped by more than 70% in a 5-year period. The price drop would have been even greater had there been a strong competing display technology. But the 19-inch color LCD, the best contender, is not yet available to challenge the CRT's reign.

4.5.2 Color CRT Number of Pixels/Inch

On Table 4.13, the model results and the actual market data on the number of pixels per inch are listed in the second and third columns. (Note that the number of pixels displayed on the screen is equivalent to the display resolution.) The actual data ranges from 76 to 135 pixels per inch, while the display model results were tuned to increase the number of pixels per inch from 54 to 93 over 6 years, an increase of nearly 7 pixels per inch/per year. The

19-inch Color CRTs		
1	2	3
Year	Price/Megapixel (Model Results) (\$/MP)	Price/Megapixel (Actual Data) (\$/MP)
1985	8035	8000
1986	6391	6000
1987	5110	2900
1988	4316	2380-3800
1989	3467	2400-5000
1990	2795	n.a.
1991	2261	2300

Table 4.12: Model results and actual market data on the price/megapixel of a 19-inch color CRT display. Sources: [1, 2, 16, 28, 38, 59, 75, 81, 83, 93].

19-inch Color CRTs		
1	2	3
Year	#Pixels/inch (Model Results)	#Pixels/inch (Actual Data)
1985	54	76
1986	59	76
1987	65	84
1988	71	67
1989	78	76-135
1990	85	n.a.
1991	93	84

Table 4.13: Model results and actual market data on the number of pixels/inch of a color CRT. Sources: [1, 2, 16, 28, 38, 59, 75, 81, 83, 93].

model was not tuned to reach the 135 pixels per inch value because the display manufacturers have no threat of competition in the resolution race, and the 135 pixels per inch value does not reflect the 1989 or 1991 color CRT resolutions available on the market. (Manufacturers can afford to increase the number of pixels per inch at a slower pace than their manufacturing laboratories can achieve.)

4.6 UNIX: Model Results and Actual Market Data

Most activities in UNIX today are performed on porting or enhancing its functionality to make it compatible with all the available hardware platforms. The following section presents the model results and the actual market data on the UNIX porting and development-from-scratch time periods and costs in Subsections 4.6.1 and 4.6.2, respectively.

4.6.1 UNIX Porting Times and Costs

Table 4.14 presents the model results and the actual market data on the UNIX porting time periods and costs. Unfortunately, the market data was only available for the years 1980 and 1991. Nevertheless, Table 4.14 presents the model results for the 11-year study period. The results display a decreasing trend, and both the porting time period and the corresponding cost results match the available data within 1%. Of course, the model was tuned to match the actual data. The model cost results were not marked up by 200% because the development-from-scratch or the porting costs are incurred by the software manufacturer or developer while developing the software. The final product

UNIX: Porting				
1	2	3	4	5
Year	Porting Time (Model Results) (Yrs)	Porting Time (Actual Data) (Yrs)	Porting Cost (Model Results) (\$)	Porting Cost (Actual Data) (\$)
1980	1.50	1.5	210000	210000
1981	1.46	n.a.	198979	n.a.
1982	1.40	n.a.	185045	n.a.
1983	1.37	n.a.	175351	n.a.
1984	1.31	n.a.	163410	n.a.
1985	1.26	n.a.	151755	n.a.
1986	1.21	n.a.	141659	n.a.
1987	1.16	n.a.	132415	n.a.
1988	1.12	n.a.	123712	n.a.
1989	1.08	n.a.	115775	n.a.
1990	1.03	n.a.	107250	n.a.
1991	0.99	1	100553	100000

Table 4.14: Model results and actual market data on the UNIX porting time periods and costs. Source: [72].

price will depend on the hardware platform on which it will run and on the software's market demand [72].

4.6.2 UNIX Development-from-Scratch Time and Cost

UNIX development-from-scratch is not as prevalent today as it was in the early 1980s, when every UNIX-based workstation manufacturer developed from scratch its own version of UNIX for its hardware platform. UNIX development-from-scratch took about 15 man-years in 1980 and about 10 man-years in 1991. Table 4.15 presents the model results and the actual market data on the UNIX development-from-scratch time periods and costs. Again, the market data was only available for the years 1980 and 1991; nevertheless, Table 4.15 presents the model results for the 11-year study period. The results display a decreasing trend, and both the development-from-scratch time period and cost results match the available data within 1%. Here, too, the model was tuned to match the actual data.

4.7 Workstation Assembly Model: Inputs and Projected Results

The inputs of the workstation assembly process model are:

- The study period starts in 1991 and ends in 1996. The year 1991 can be used to compare the model's results with actual workstation prices. Because of rapid technological changes in the industry, a period of 5 years is sufficiently long, given relatively short product lives and innovation uncertainties.

UNIX: Development-from-Scratch				
1	2	3	4	5
Year	Development Time (Model Results) (Yrs)	Development Time (Actual Data) (Yrs)	Development Cost (Model Results) (\$)	Development Cost (Actual Data) (\$)
1980	15.00	15	2100000	2100000
1981	14.64	n.a.	1989785	n.a.
1982	14.02	n.a.	1850450	n.a.
1983	13.69	n.a.	1753510	n.a.
1984	13.14	n.a.	1634101	n.a.
1985	12.57	n.a.	1517553	n.a.
1986	12.08	n.a.	1416595	n.a.
1987	11.63	n.a.	1324151	n.a.
1988	11.19	n.a.	1237116	n.a.
1989	10.79	n.a.	1157750	n.a.
1990	10.30	n.a.	1072502	n.a.
1991	9.94	10	1005527	1000000

Table 4.15: Model results and actual market data on the UNIX development-from-scratch time periods and costs. Source: [72].

- The component supply simulation models presented in earlier sections of this chapter were used to project the price per megaHertz of CPUs, the price per megabyte of DRAMs, the price per megabyte of magnetic hard disk, and the price of a 19-inch color CRT from 1991 to 1996. It is assumed that the component supply models' input parameters and the rates of change of each of the components' physical characteristics trends listed in Section 4.2 hold for the 1980-1996 period. These projected results⁸ and other assumptions are listed on Table 4.16. The price of the operating system is set at \$800 [86], and the prices of each electric power supply, mounting board, cable, mouse, and keyboard are set at \$100, \$20, \$50, \$50, and \$100, respectively [33]. Since all the components' costs have been marked up by 200% in Chapter 4, no additional workstations price markups are included in the model.

Each workstation has a vector of attributes, and these attributes determine the price of the workstation. For the sake of comparison, three types⁹ of workstations are considered, and it is assumed that the manufacturer has the capabilities to produce all of them:

- Type 1 (Economy Model): 50 MHz CPU, 16 megabytes of DRAM, 250 megabytes of magnetic storage, a 19-inch color CRT monitor, and a UNIX operating system with the documentation.

⁸The projection years are actually 1992 to 1996 because the 1991 results can be validated with actual market data.

⁹The workstation type refers to the vector of workstation *attributes*: the speed of the processor, the amount of main memory, the capacity of the hard disk, the resolution of the display, and the software programs loaded into the machine's hard disk.

Process Model						
	1991	1992	1993	1994	1995	1996
CPU(\$/MHz)	7.1	6.8	6.5	3.08	2.9	2.7
DRAM(\$/MB)	79	60	35	25	20	14
Disk(\$/MB)	6.1	5	4.1	3.3	2.7	2.2
UNIX (each)	800	800	800	800	800	800
Display (each)	3416	3322	3232	3146	3064	2985
Power Supply (each)	100	100	100	100	100	100
Board (each)	20	20	20	20	20	20
Cable (each)	50	50	50	50	50	50
Mouse (each)	50	50	50	50	50	50
Keyboard (each)	100	100	100	100	100	100

Table 4.16: Projected prices of the workstation assembly components and raw materials. Sources: [33, 86].

- Type 2 (Mid-Range Model): 100 MHz CPU, 32 megabytes of DRAM, 500 megabytes of magnetic storage, a 19-inch color CRT monitor, and a UNIX operating system with the documentation.
- Type 3 (Top-of-the-Line Model): 200 MHz CPU, 64 megabytes of DRAM, 1000 megabytes of magnetic storage, a 19-inch color CRT monitor, and a UNIX operating system with the documentation.

4.7.1 Projected Results

Figure 4.1 presents a log-linear graph that shows how the present value prices of assembled workstation types 1, 2, and 3 decrease over the 1991-1996 period. The objective function of the linear process model (Equation 3.92) was minimized for each of the proposed workstation types, and the base 10 logarithm of their present value prices are plotted in Figure 4.1 for each year of the 1991-1996 study period.

Figure 4.1 shows that the present value prices of all workstation types decrease by more than 65% in 5 years. This is a rate of decrease of over 20% per year for fixed capabilities workstations. (In 1996, the price of a 19-inch color CRT display is close to 60% of the total price of an Economy Model workstation—type 1.) These results do not reflect actual workstation prices. Their sole purpose is to show the correlation between the price decrease of the assembled workstation and its components.

The price of the top-of-the-line workstation will be less than \$10,000

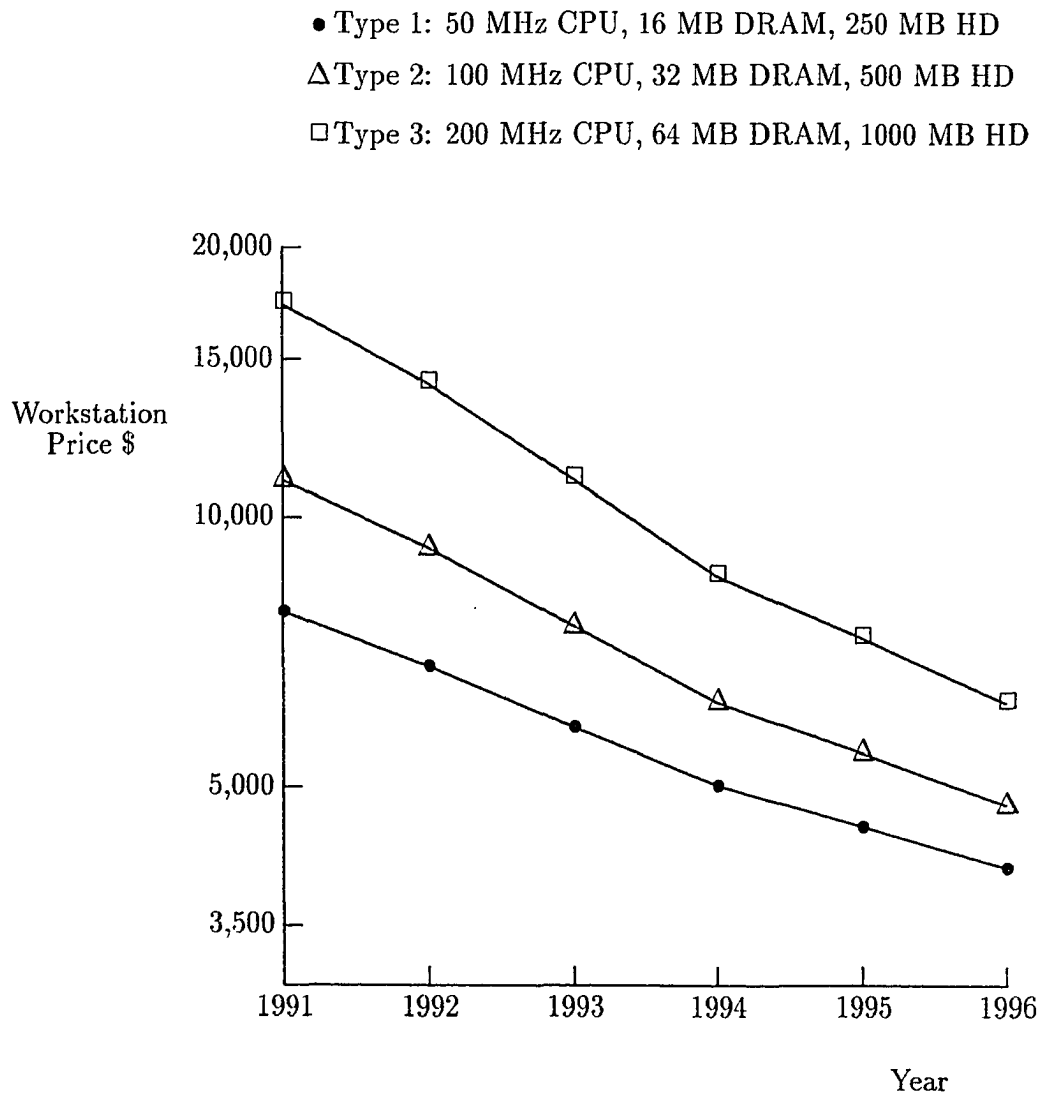


Figure 4.1: Present value prices of workstations: Types 1, 2, and 3.

by 1994. Not everyone will be happy about this. For example, Cray Research¹⁰ sounded a warning in 1991 [58], acknowledging that the most serious threat to its business is the emergence of the superworkstation which, for less than \$10,000, will be capable of performing the same computations that a supercomputer¹¹ does now—and in only twice the time!

Remark. Since the workstation assembly model did not include plant expansion costs, or import, export, and government quota restrictions, and it considered only one assembly plant and one market in its vicinity—i.e., no transportation costs were involved—its formulation can forgo cost minimization in the objective function (Equation 3.92). The costs of the assembled workstations can be computed by multiplying the configuration requirements of each workstation type by the corresponding component prices on Table 4.16. Nevertheless, the presented assembly model was formulated as such to provide a foundation upon which future research can expand.

4.8 Sensitivity Analyses

To mainframe and supercomputer manufacturers, the emergence of the workstation has meant cuts in market shares and profit margins [58]. In response, mainframe manufacturers have transferred their sophisticated technologies to the workstation platform and introduced high end workstations which can outperform their original mainframes. HP announced in 1991 a high-end work-

¹⁰Cray Research is the manufacturer of the fastest supercomputers in the world, to date.

¹¹A typical supercomputer costs \$2 to \$3 million.

station which is based on its mainframe's RISC Precision Architecture platform, and IBM announced in 1990 a high-end workstation which is based on its mainframe's superscalar RISC architecture. All of these technology migrations have become possible because manufacturers have achieved high production yields as the technology-driving trends of their hardware improve.

What will shape the trends of the future? This is impossible to answer with certainty. We think, however, that models of the type developed in this dissertation can help us understand where the industry might be going. To illustrate, this section presents sensitivity analyses of how the attributes of assembled workstations might change in response to changes in certain technology-driving trends. The discrete event simulation supply models and the linear workstation assembly process model provided in Chapter 3 of this dissertation are used to perform these analyses. Each change in a technology-driving trend is represented with a particular case number. In turn, the results of each change case are compared with the Base Case results for the period of analysis.

This section examines the models' behavior in response to the following two "*What if...?*" questions:

1. How sensitive are the prices of workstation components and assembled workstations to a faster rate of decrease and to no decrease in the IC's **feature size (FS)**?
2. How sensitive are the IC die yields and the CPU and DRAM prices to an increase and to a decrease in the **number of silicon wafer defects per unit area (DPUA)**?

The results are presented in the following subsections:

- Subsection 4.8.1 studies the effects of the feature size (FS) of ICs on the prices of the components and, ultimately, on the prices of the assembled workstations. The special case wherein the price per megabyte of DRAM becomes cheaper than the magnetic hard disk's is also discussed. (The effects of the feature size on the pricing of one copy of the UNIX operating system are not presented in this section since only the UNIX development-from-scratch and porting costs are modeled in Section 3.5. In general, the price of the operating system is equal to a fixed markup to the overall workstation cost set in this dissertation at \$800 [86].)
- Subsection 4.8.2 studies the effects of the silicon wafer defects per unit area (DPUA) on the die yields and, consequently, on the price per megahertz of CPUs and the price per megabyte of DRAMs.

4.8.1 Sensitivity to the Feature Size

The feature size is one of the main technology-driving trends of the VLSI technologies. It has been estimated that the feature size will not run into any physical barriers until the end of the century due to the recent achieved improvements in lithography and silicon etching processes [21].

However, there is concern among many in the electronics industry that the feature size might reach a physical barrier and stop decreasing¹² in the next

¹²A physical barrier might only slow down the decrease in the feature size. Since this is a "What if...?" illustration and there is no way to predict with certainty the consequences of the physical barrier, only the extreme case—stoppage of feature size decrease—was considered.

decade [18, 21]. To study the effects of such an event, the models developed in Chapter 3 were simulated with a stoppage in decreasing feature size after 1992. The year 1992 was chosen instead of the year 2000, for example, because of the rapid pace at which the computer industry is changing and the uncertainties associated with it; a later year may be too far into the future for us to analyze the consequences of a stoppage in decreasing feature size. Again, feature size sensitivity analyses are meant to shed some light on how important a continually decreasing feature size is to the electronics industry and the computer industry in general. To make the study complete, the models were also simulated with an accelerated decrease of the feature size after 1992.

The alternative feature size trends were implemented as follows:

- **Case 1 - Base Case**

The feature size follows the trend presented in Equation 3.14:

$$FS_t = FS_{BC} = 10^{1.4-0.055*(t-1960)} \text{ (micron)}. \quad (4.3)$$

In this case the feature size continually decreases at 5.5% per year for the period of study.

- **Case 2 - $FS_t = FS_{BC/1992}$ for $t > 1992$**

The feature size stops decreasing after 1992; hence, for $t > 1992$, the feature size values are equal to the Base Case's value in 1992:

$$FS_t = FS_{BC/1992} = 10^{1.4-0.055*(1992-1960)} \text{ (micron) for } t > 1992. \quad (4.4)$$

- **Case 3 - $FS_t < FS_{BC}$ for $t > 1992$**

The feature size decreases faster than expected after 1992; hence, for $t >$

1992, the feature size values are equal to the Base Case's value in 1992 multiplied by the new rate of decrease of 10%:

$$FS_t = FS_{BC/1992} * 10^{-0.1*(t-1992)} \text{ (micron) for } t > 1992. \quad (4.5)$$

For each of the cases presented above, the ICs, magnetic hard disk, and color CRT display supply models of Chapter 3 were simulated for the 1992-1996 period. It is assumed that the component supply models' input parameters and the rates of change of each of the components' physical characteristics trends listed in Section 4.2 hold for the 1980-1996 period.

The organization of this subsection is as follows:

- The first four paragraphs illustrate the sensitivity to changes in the feature size of the price per megaHertz of CPUs, the price per megabyte of DRAMs, the price per megabyte of magnetic hard disks, and the prices of a 19-inch color CRT for each of the three cases presented above.
- The fifth paragraph illustrates the collective effect of the changes in the feature size on the assembled workstation prices for each of the three cases presented above.
- The sixth paragraph illustrates how, if feature size decreases at a faster rate than the Base Case's, the price per megabyte of DRAMs will become lower than the price per megabyte of magnetic storage in the early part of the next decade.

Sensitivity of the CPU Price/MHz to Feature Size. Figure 4.2 illustrates in a log-linear graph the decreasing price per megaHertz of CPUs for the 1992-1996 period. The ICs supply model was simulated to compute for each of the three cases the prices of CPUs with three different speeds—50, 100, and 200 MHz. The prices of the CPUs were then divided by their corresponding speeds to compute the price per megaHertz values. The base 10 logarithm of the lowest price per megaHertz values are plotted in Figure 4.2 for each year of the 1992-1996 period.

It is evident that, when the feature size stops decreasing, the price per megaHertz values are higher than the Base Case's. Furthermore, the price per megaHertz values are lower than the Base Case values where there exists an accelerated decrease in feature size. By almost doubling the rate of decrease of the feature size, from 5.5% to 10%, and by stopping its decrease after 1992, the price per megaHertz values differ from the Base Case values by as much as 7.4% and 129.6% respectively in 1996. However, the most noticeable characteristic of Figure 4.2 is the drop in the price per megaHertz in the Base Case and Case 3 values after 1993 and 1992. The reason for those two drops is the packaging costs of the CPUs. As the feature size decreases, the die area of a fixed speed CPU gets smaller, and as the die area gets smaller than 1.1 cm^2 , the packaging cost drops by \$47 (refer to Equations 3.34 and 3.35). Hence, the drops in the graphs of Figure 4.2. The Case 3 graph drops earlier because the feature size is decreasing faster than in the Base Case.

Sensitivity of the DRAM Price/MB to Feature Size. Figure 4.3 illustrates in a log-linear graph the decreasing price per megabyte of DRAMs for the

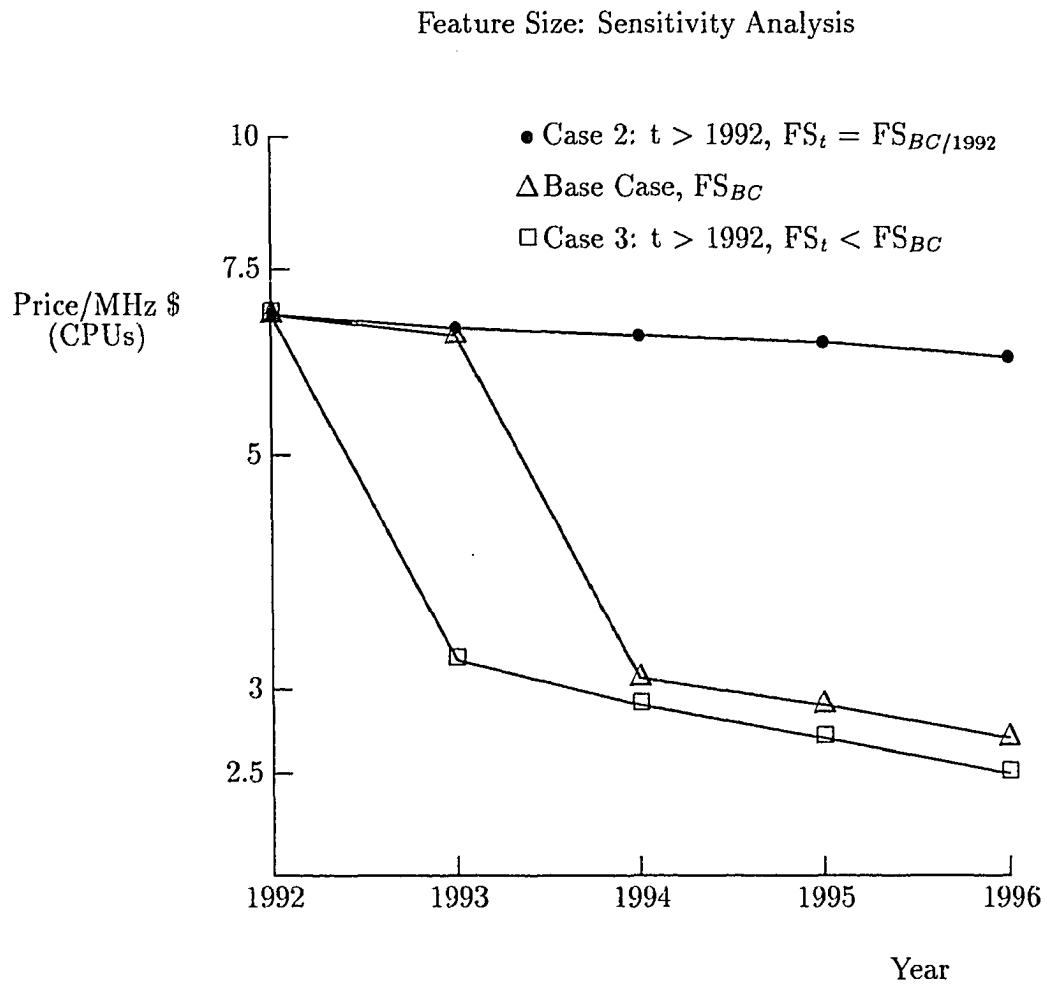


Figure 4.2: Feature Size: Sensitivity of the price/megaHertz of CPUs - Cases 1, 2, and 3.

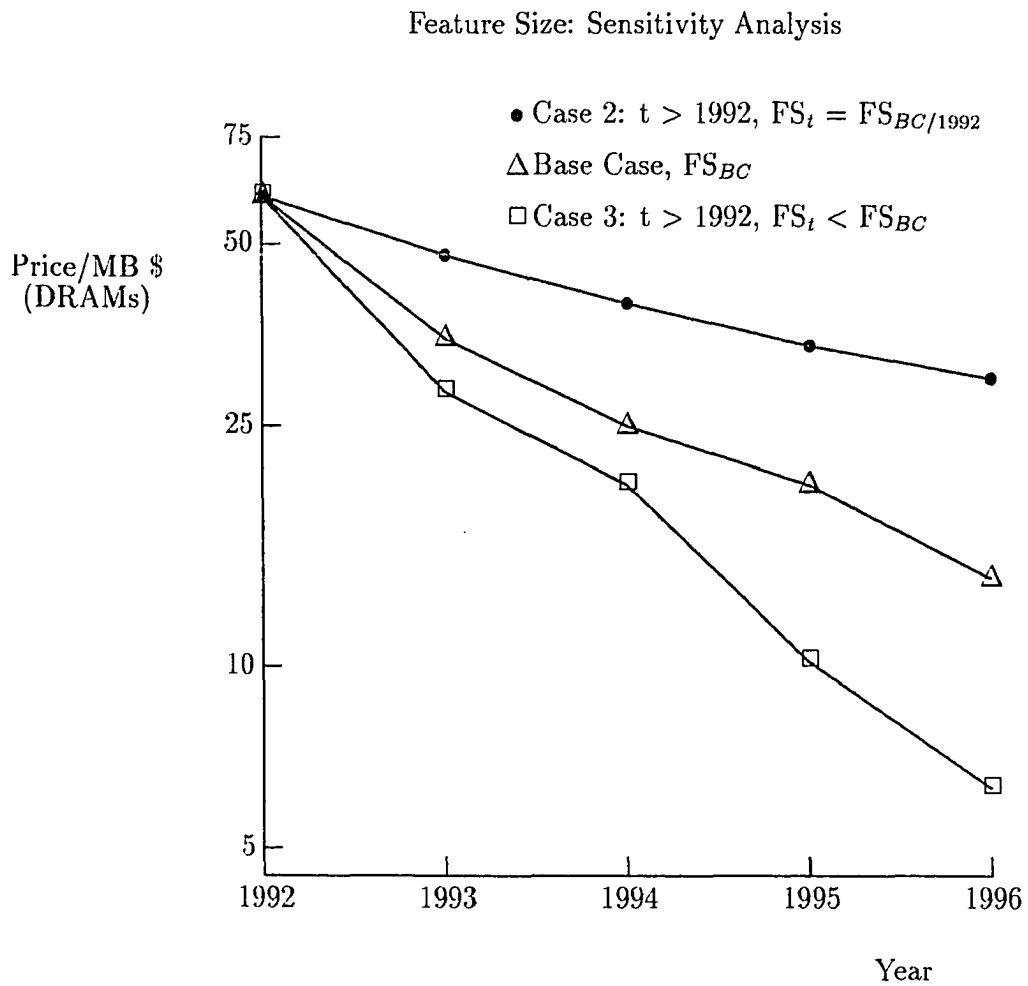


Figure 4.3: Feature Size: Sensitivity of the price/megabyte of DRAMs - Cases 1, 2, and 3.

1992-1996 period. The ICs supply model was simulated to compute, for each of the three cases, the prices of DRAMs with six different capacities—1, 2, 8, 32, 64, and 128 megabytes. The prices of the DRAMs were then divided by their corresponding capacities to compute the price per megabyte values. The base 10 logarithm of the lowest DRAM price per megabyte values are plotted in Figure 4.3 for each year of the 1992-1996 period.

It can be seen in Figure 4.3 that, when the feature size stops decreasing, the DRAM price per megabyte values are higher than the Base Case's. Moreover, the DRAM price per megabyte values are lower than the Base Case values for an accelerated decrease in feature size. By almost doubling the rate of decrease of the feature size, from 5.5% to 10%, and by stopping its decrease after 1992, the price per megabyte values of DRAMs differ from the Base Case values by as much as 55% and 114.3% respectively in 1996. In response to the drop in the packaging costs, the drops in the price per megabyte in Figure 4.3 are not as pronounced as those in Figure 4.2. Nonetheless, they do occur in 1995 and 1994 for the Base Case and Case 3, respectively.

Sensitivity of the Magnetic Hard Disk Price/MB to Feature Size.

Figure 4.4 illustrates in a log-linear graph the decreasing price per megabyte of magnetic hard disks for the 1992-1996 period. The magnetic hard disk supply model was simulated to compute for each of the three cases the hard disk price per megabyte values. The feature size affects the performance of the channel's microcontroller (Equation 3.64), which in turn affects the price per megabyte values of magnetic hard disks (Equation 3.61); (refer to the magnetic hard disk supply model of Section 3.3 for more details).

Feature Size: Sensitivity Analysis

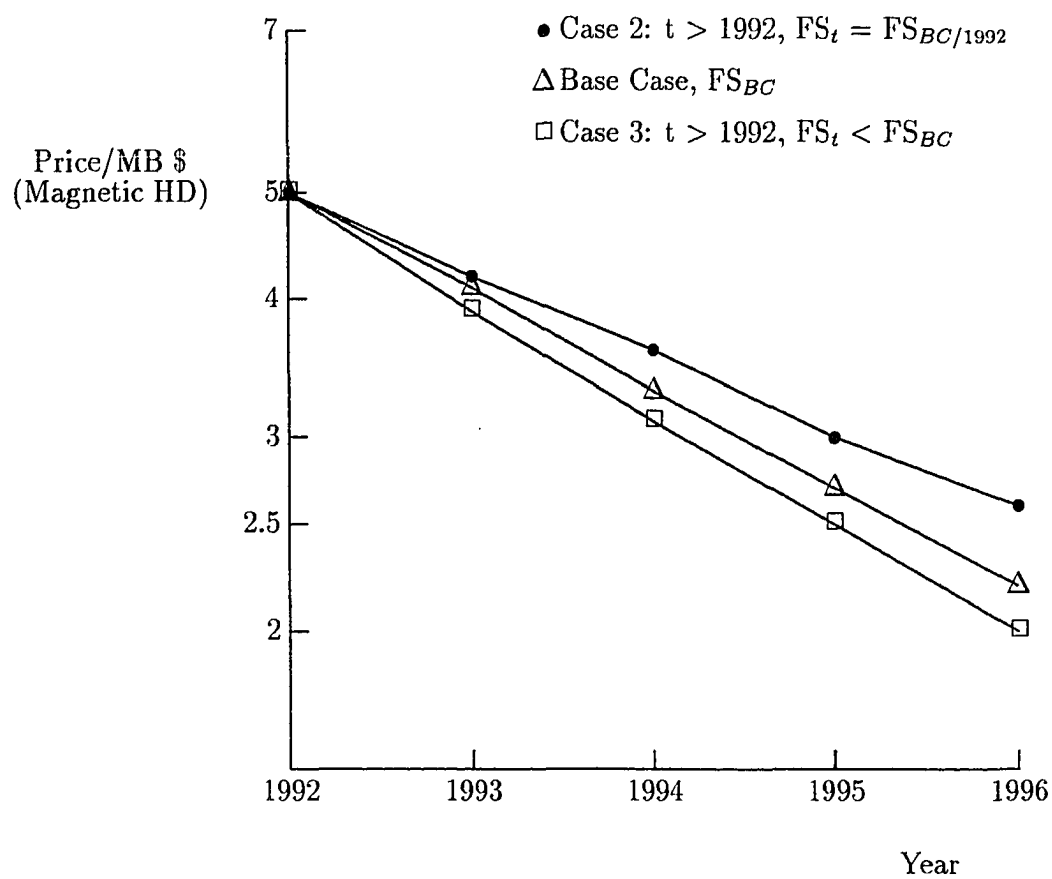


Figure 4.4: Feature Size: Sensitivity of the price/megabyte of magnetic hard disks - Cases 1, 2, and 3.

It can be seen in Figure 4.4 that when the feature size stops decreasing, the magnetic hard disk price per megabyte values are higher than the Base Case's. The figure further shows that the price per megabyte values are lower than the Base Case values when there exists an accelerated decrease in feature size. By almost doubling the rate of decrease of the feature size, from 5.5% to 10%, and by stopping its decrease after 1992, the price per megabyte values of magnetic hard disks differ from the Base Case values by as much as 9.1% and 18.2% respectively in 1996.

Sensitivity of the Color CRT Prices to Feature Size. Figure 4.5 illustrates in a log-linear graph the decreasing prices of a 19-inch color CRT display for the 1992-1996 period. The CRT display supply model was simulated to compute for each of the three cases the 19-inch color CRT prices. The feature size affects the speed and the capacity of the microprocessors and the DRAMs used in the CRT image drivers. Consequently, the feature size affects the number of pixels per inch displayed on the screen (Equation 3.76), the screen's resolution (Equation 3.70), and the cost per megapixel (Equation 3.80).

It can be seen in Figure 4.5 that, when the feature size stops decreasing, the 19-inch color CRT display prices are higher than the Base Case's. Moreover, the 19-inch color CRT prices are lower than the Base Case prices where there exists an accelerated decrease in feature size. By almost doubling the rate of decrease of the feature size, from 5.5% to 10%, and by stopping its decrease after 1992, the 19-inch color CRT prices differ from the Base Case prices by as much as 4.8% and 6.3% respectively in 1996.

Feature Size: Sensitivity Analysis

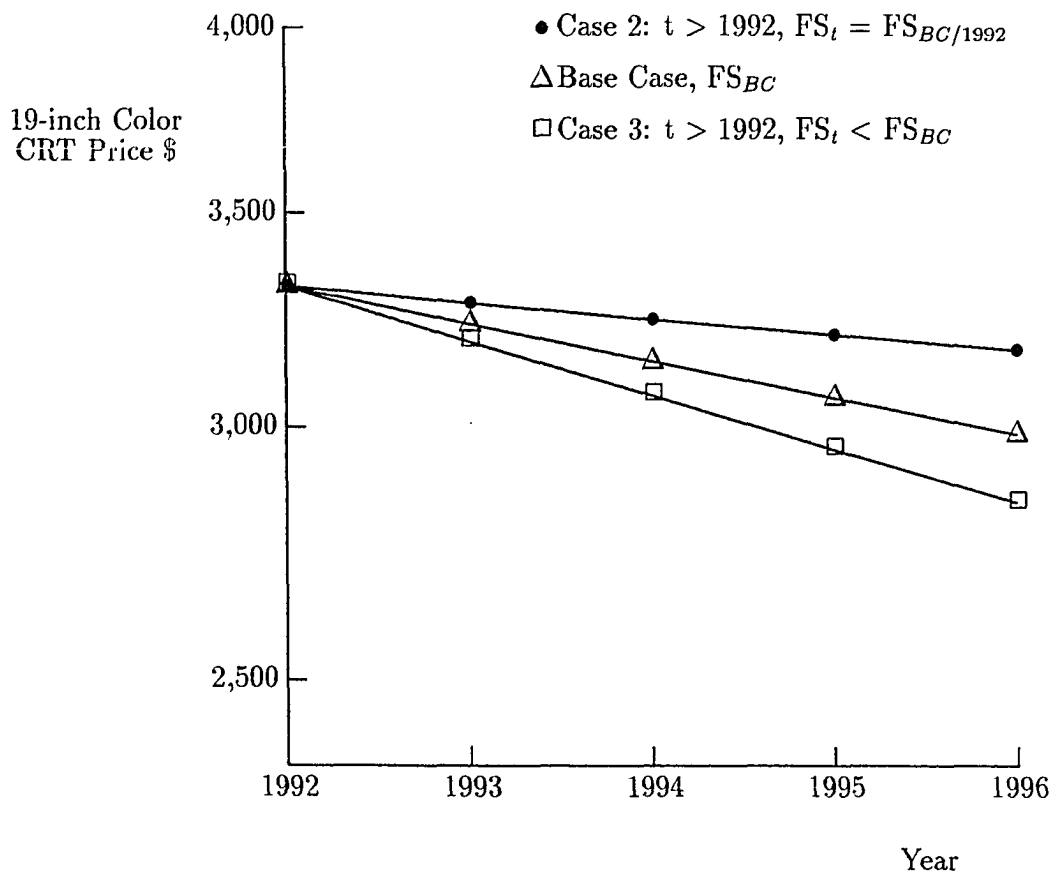


Figure 4.5: Feature Size: Sensitivity of the price of a 19-inch color CRT display - Cases 1, 2, and 3.

Sensitivity of the Assembled Workstations Prices to Feature Size. As has been shown, the feature size affects the attributes and prices of the workstation hardware components. In turn, it also affects the price of the workstation. The workstation assembly process model in Section 3.6 was used to illustrate the effects of changes in the feature size on the price of an assembled workstation with a 100 MHz CPU set, a 32 MB main memory, a 500 MB magnetic hard disk, and a 19-inch color CRT. Figure 4.6 presents the magnitude of the present value of the workstation prices for the three feature size cases.

It can be seen that when the feature size stops decreasing, the workstation prices are higher than the Base Case prices. The figure also shows that the workstation prices are lower than the Base Case prices for a faster decreasing feature size. By almost doubling the rate of decrease of the feature size, from 5.5% to 10%, and by stopping its decrease after 1992, the workstation prices differ from the Base Case prices by as much as 8.4% and 20.5% respectively in 1996.

Feature Size: Sensitivity of the Price/MB - DRAMs versus Magnetic Storage. One of the main objectives of the Sematech consortium was to develop denser DRAM ICs and help the US regain some of its lost DRAM market share from the Japanese manufacturers. But the good news for the US ICs manufacturers is a source of worry for the magnetic hard disk manufacturers. Not only do they need to fend off the competition from the optical technology, but denser DRAMs might have a lower price per megabyte, and if nonvolatile¹³

¹³A nonvolatile DRAM retains the data stored in it even when the power is switched off.

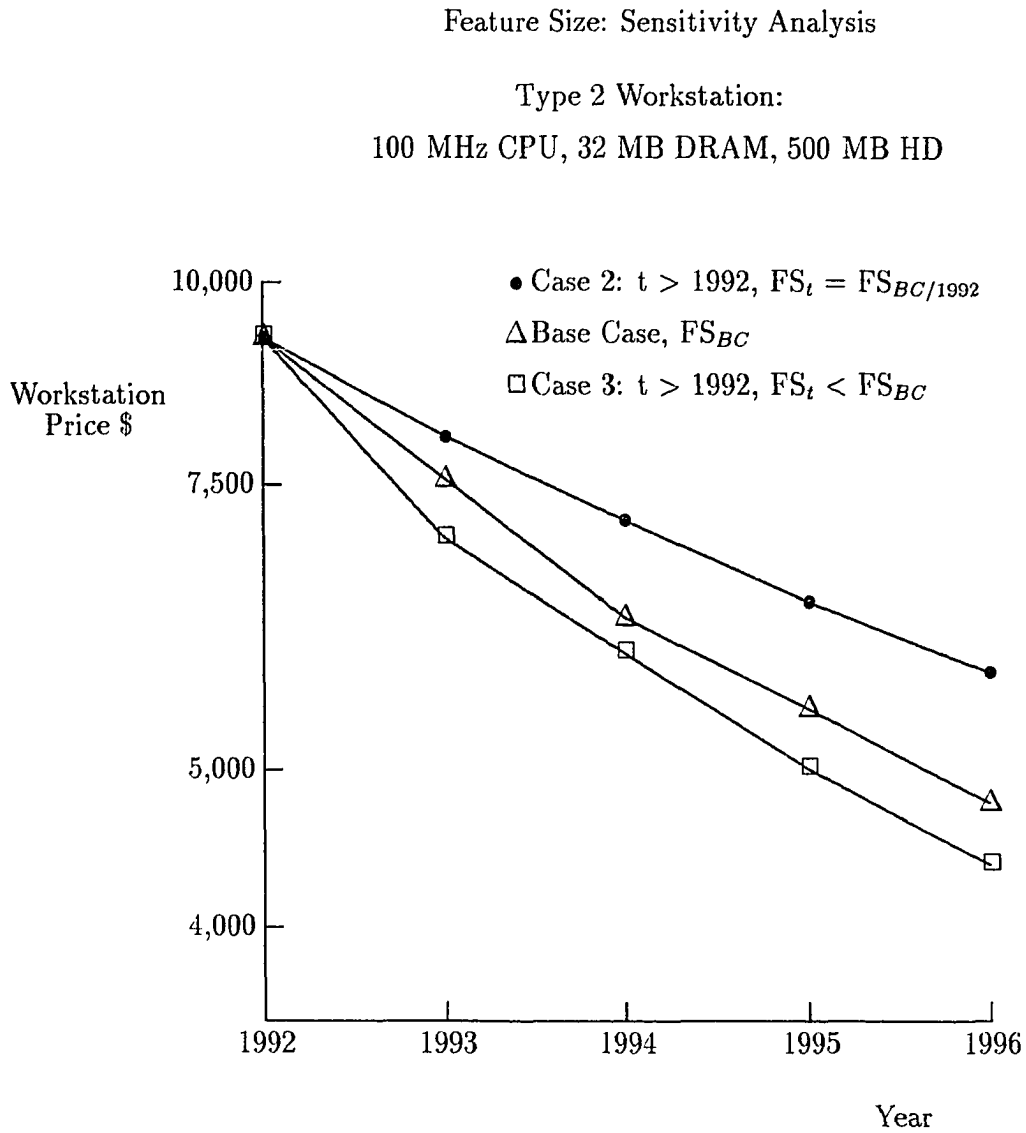


Figure 4.6: Feature Size: Sensitivity of the present value price of a type 2 workstation - Cases 1, 2, and 3.

ones are developed, DRAMs might replace the magnetic hard disk as the main computer storage component.

In this paragraph, a scenario is analyzed which results in the DRAM price per megabyte becoming cheaper than magnetic storage. If the ICs and magnetic results presented in Chapter 4 are projected to the year 2000, the DRAM price per megabyte will not catch up with magnetic storage during this decade. However, if the feature size decreases at a faster rate (Case 3) than was projected in the Base Case, and the magnetic storage manufacturers are not able to improve their products' attributes at a faster rate than was presented in Section 3.3, the DRAM price per megabyte might become cheaper than magnetic storage during the first half of the next decade. What follows is an illustration of this phenomenon.

Only the Base Case and Case 3 of the feature size trends are considered for this analysis. For each one of the cases, the ICs and magnetic storage supply models of Chapter 3 were simulated for the 1992-2005 period and the results are presented in Figure 4.7. It is assumed that the ICs and magnetic storage supply models' input parameters and the rates of change of each of the components' physical characteristics trends listed in Section 4.2 hold for the 1980-2005 period.

Figure 4.7 illustrates in a log-linear graph the decreasing prices per megabyte of DRAMs and magnetic storage for the 1992-2005 period. The ICs and magnetic storage supply models were simulated to compute for each of the two cases the prices of DRAMs with seven different capacities—1, 2, 8, 32, 64, 128, and 256 megabytes—and the price per megabyte of magnetic storage. The prices of the DRAMs were then divided by their corresponding capacities to

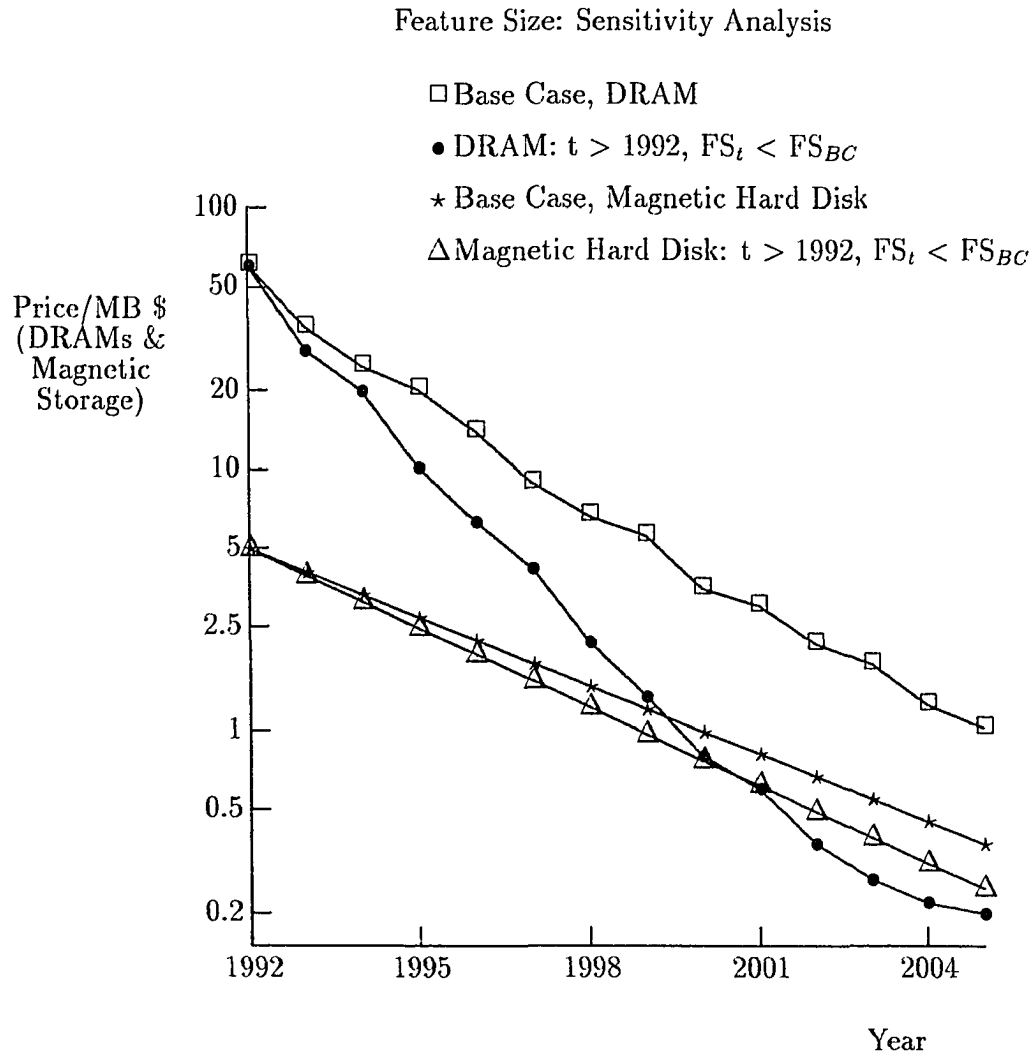


Figure 4.7: Feature Size: Sensitivity of the DRAM and the magnetic hard disk prices/megabyte - Cases 1 and 3.

compute the price per megabyte values. The base 10 logarithm of the lowest DRAM price per megabyte values are plotted in the upper two graphs of Figure 4.7 for each year of the 1992-2005 period, and the magnetic storage price per megabyte values are plotted in the lower two graphs for each of the two cases. As can be seen in Figure 4.7, the DRAM price per megabyte becomes less than the magnetic storage price per megabyte around 2001.

The fact that the magnetic storage price per megabyte values do not decrease faster than projected can be attributed to the following. As the feature size decreases faster than expected, the microcontroller in the hard disk's channel can operate faster (see Equation 3.22), the price per microcontroller MIPS decreases (see Section 4.3), and, consequently, the price per megabyte values of magnetic storage become less than the Base Case's (see Equations 3.61 and 3.64). However, the channel's performance improvement does not affect the price per megabyte of magnetic storage enough to decrease it faster than the DRAM's. This is because the net decrease of the DRAM price per megabyte values is approximately 2 over the period of study (see Equations 3.25 and 3.9), while the net decrease in value of the magnetic storage price per megabyte values is only about 0.3 (see Equations 3.22, 3.64, and 3.9), yielding, approximately, a 7:1 ratio of difference in their sensitivity to the feature size.

4.8.2 Sensitivity to the Number of Silicon Wafer Defects per Unit Area

The die yield plays a crucial role in determining the price of an IC and whether the IC ought to remain in the production line or be removed. As

the production process of ICs gets more and more complicated, precision and defects-free production become harder to achieve. One of the main steps in producing defect-free and cheaper ICs is to reduce the number of silicon wafer defects per unit area (DPUA). A defect can occur anytime during the production process of the wafer itself or the IC.

Clean room technologies and learning have been two key factors in improving the yields of good manufactured dies. Since DPUA reflects how clean, efficient, and accurate the manufacturing of an ICs plant is, its effect on the ICs prices will be studied in this subsection, in particular, the price per megaHertz of CPUs and the price per megabyte of DRAMs. DPUA was assumed to remain constant at 2.5 defects per cm^2 during the 1980-1991 simulation period. To study the effects of varying DPUA, the ICs supply model was simulated from 1992 to 1996, with DPUA set in one case to twice its original value—5 defects/ cm^2 —and in a second case to half its original value—1.25 defects/ cm^2 —after 1992. It is assumed that the ICs supply models' input parameters and the rates of change of each of the ICs' physical characteristics trends listed in Section 4.2 hold for the 1980-1996 period.

What follows are the 3 cases used to show the effects of changes in the number of silicon wafer defects per unit area on the IC die yields (the first paragraph), the price per megaHertz of CPUs (the second paragraph), and the price per megabyte of DRAMs (the third paragraph):

- **Case 1 - Base Case**

$$DPUA_t = DPUA_{BC} = 2.5 \text{ defects/cm}^2 \text{ for } t > 1992. \quad (4.6)$$

- **Case 2**

$$DPUA_t = 2 * DPUA_{BC} = 5 \text{ defects/cm}^2 \text{ for } t > 1992. \quad (4.7)$$

- **Case 3**

$$DPUA_t = \frac{1}{2} * DPUA_{BC} = 1.25 \text{ defects/cm}^2 \text{ for } t > 1992. \quad (4.8)$$

DPUA: Sensitivity of the IC Die Yields. The first factor reflecting the change in the value of $DPUA_{BC}$ is the die yield. Equation 4.9 shows that the higher (or the lower) the DPUA number is, the lower (or the higher) the manufacturing die yield:

$$DY_t = WY_t * \left(1 + \frac{DPUA_t * DA_t}{CM_t}\right)^{-CM_t}. \quad (4.9)$$

Table 4.17 lists in the columns labeled Case 2, Base Case¹⁴, and Case 3 the die yield model results for these cases. When the $DPUA_{BC}$ increases by 100% to 5 defects per cm^2 , the die yield values in Column 2 decrease by as much as 88.9%. Similarly, as $DPUA_{BC}$ decreases by 50% to 1.25 defects per cm^2 , the die yield values in Column 4 increase by as much as 355.6%, or almost 7 times the decrease in the Base Case DPUA value. The die yield results for the three cases reflect how sensitive the die yield is to clean and precise production environments and processes. A 10% decrease in the DPUA could result in a 70% improvement in the die yield and, consequently, a sizable decrease in the die prices over time.

¹⁴The Base Case die yield model results are taken from Table 4.1.

Sensitivity Analysis			
	Case 2	Base Case	Case 3
Die Area (cm ²)	Die Yield (%) DPUA = 5	Die Yield (%) DPUA = 2.5	Die Yield (%) DPUA = 1.25
0.0625	66.8	77.3	88.3
0.2601	30.2	49.8	66.1
0.5776	11.5	27.3	47
1.0404	4.1	13.4	30.2
1.6129	1.6	6.6	18.8
2.3104	0.7	3.3	11.5
3.1684	0.3	1.6	6.8
4.1209	0.1	0.9	4.1

Table 4.17: DPUA: Sensitivity of the die yields for certain die areas - Cases 1, 2, and 3.

DPUA: Sensitivity of the CPU Price/MHz. Figure 4.8 illustrates in a log-linear graph the decreasing price per megaHertz of CPUs for the 1992-1996 period. The ICs supply model was simulated to compute for each of the three cases the prices of CPUs with three different speeds—50, 100, and 200 MHz; the prices of the CPUs were then divided by their corresponding speeds to compute the price per megaHertz values. The base 10 logarithm of the lowest price per megaHertz values are plotted in Figure 4.8 for each year of the 1992-1996 period.

It can be seen in Figure 4.8 that, when the $DPUA_{BC}$ is doubled to 5 defects per cm^2 , the price per megaHertz value jumps to more than double the Base Case value in 1993, reaching a 277% difference in 1996. Moreover, as the $DPUA_{BC}$ is halved to 1.25 defects per cm^2 , the price per megaHertz value drops to less than half the Base Case value in 1993, reaching a 51.8% difference in 1996. (Note also the effects of the change in packaging costs in the three graph drops in 1993, which were discussed earlier.)

DPUA: Sensitivity of the DRAM Price/MB. Figure 4.9 illustrates in a log-linear graph the decreasing price per megabyte of DRAMs for the 1992-1996 period. The ICs supply model was simulated to compute for each of the three cases the prices of DRAMs with six different capacities—1, 2, 8, 32, 64, and 128 MB; the prices of the DRAMs were then divided by their corresponding capacities to compute the price per megabyte values. The base 10 logarithm of the lowest DRAM price per megabyte values are plotted in Figure 4.9 for each year of the 1992-1996 period.

Figure 4.9 shows that, as the $DPUA_{BC}$ is doubled to 5 defects per cm^2 , the price per megabyte value jumps to more than double the Base Case value in

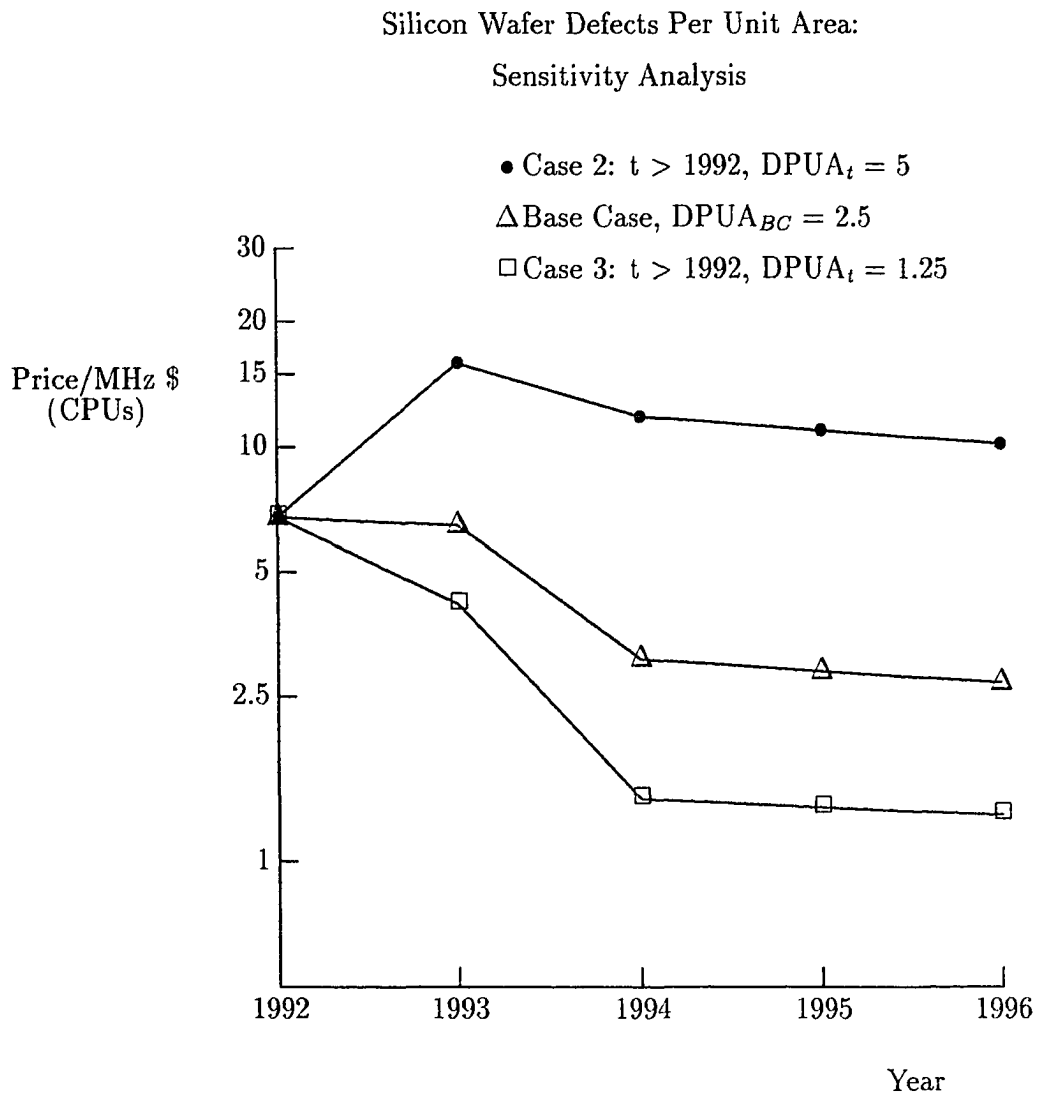


Figure 4.8: DPUA: Sensitivity of the price/megaHertz of CPUs - Cases 1, 2, and 3.

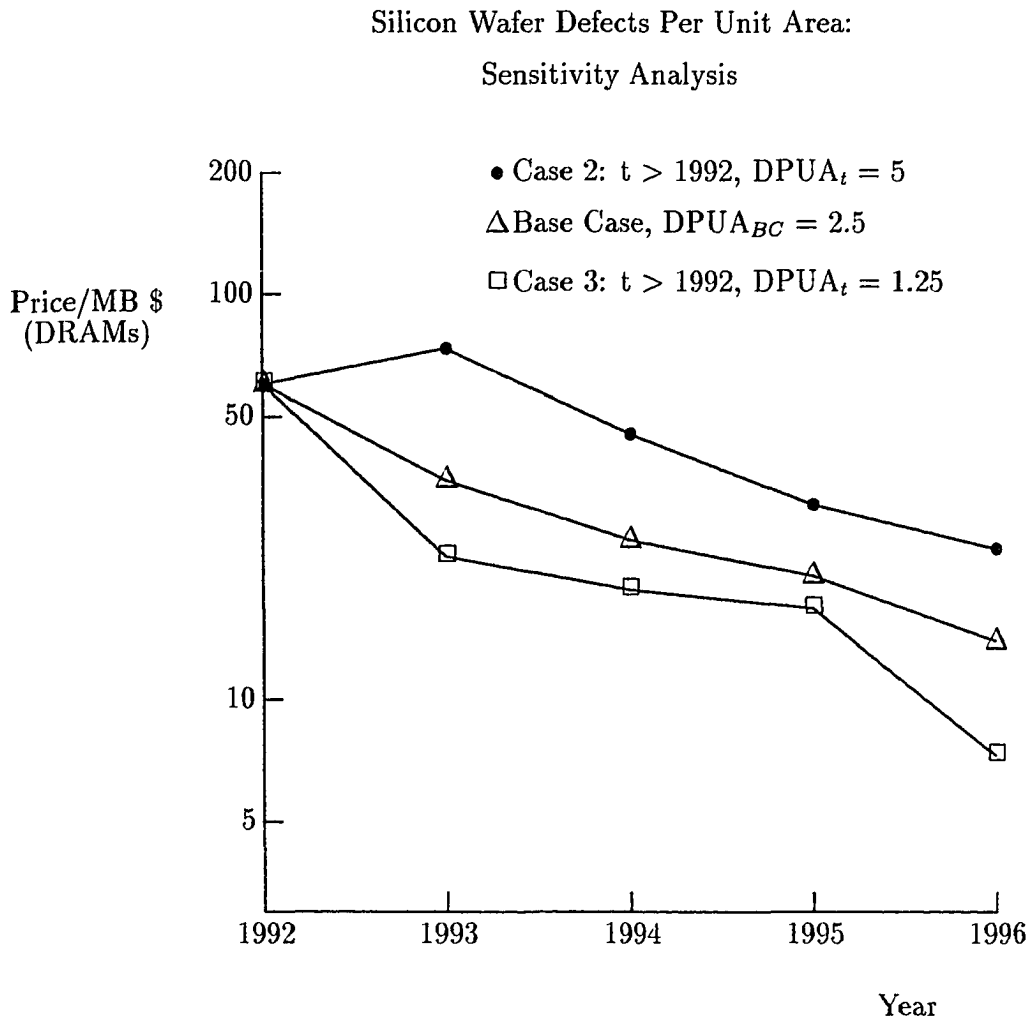


Figure 4.9: DPUA: Sensitivity of the price/megabyte of DRAMs - Cases 1, 2, and 3.

1993, climbing up to a 69.1% difference in 1996. When the $DPUA_{BC}$ is halved to 1.25 defects per cm^2 , the price per megabyte value drops to less than half the Base Case value in 1993, reaching a 47.7% difference in 1996. (Here, too, note the effects of the difference in packaging costs in the three graph drops in 1993.)

As seen in the percentage differences between the Base Case and the Case 2 and Case 3 results, the ICs die yields, the CPU price per megaHertz, and the DRAM price per megabyte results are far more sensitive to an increase in the number of silicon wafer defects per unit area than they are to a decrease.

Chapter 5

Conclusions and Suggestions for Future Research

This dissertation approached the dynamics of the computer industry through the workstation sector and its components. Discrete event simulation supply models for workstation components, including microprocessors, DRAMs, magnetic hard disks, color CRT displays, and UNIX operating systems were developed to study the effects of improving technology-driving trends on the capabilities and prices of the components and, ultimately, on the attributes of the assembled workstations. A relational diagram concept was used in Chapter 3 to communicate the model relationships of the attributes of the components over time. The supply models were simulated over the 1980-1991 study period to tune their results to match with actual market data. Since the supply models provided *cost* results only, a 200% markup was applied to the cost results to compare them with the actual market *price* data. Most of the results of the supply models and the actual market data differed by less than 10%.

To study the overall effects of decreasing prices and improving capabilities of the components on the supply of assembled workstations, a linear workstation assembly process model was developed. The objective function was to minimize the present value of the total components and raw materials costs that went into the assembly of the workstations. The price results of the component supply models were projected over the 1992-1996 period and used as input

prices in the assembly model. Three workstation types were considered, and all the results showed a decreasing trend of workstation prices in the years to come. By 1994, a top-of-the-line workstation will cost less than \$10,000 and will have the following hardware capabilities: a 200 megaHertz CPU, a 64 megabyte main memory, a 1 gigabyte magnetic hard disk, and a 19-inch color CRT display with a 2.6 megapixel resolution. These same capabilities, if configured into a workstation in 1991, would cost approximately \$20,000. This is a rate of decrease in price of over 20% per year for a fixed capabilities workstation.

Finally, sensitivity analyses were performed for the models for three different rates of decrease of the IC feature size and for three different numbers of silicon wafer defects per unit area. It was apparent that all the results of the supply models were sensitive to the changes in the feature size, and their sensitivity was echoed in the overall prices of the assembled workstations. A steady decrease in the feature size helps IC manufacturers stay competitive in a semiconductor market that has become more capital intensive and more competitive than ever. The most striking results of the feature size sensitivity analyses were the prices per megabyte of DRAMs and magnetic storage. As the feature size's rate of decrease doubled, the rates of decrease of both DRAM price per megabyte and magnetic storage price per megabyte increased; however, DRAM price per megabyte decreased at a faster rate than did magnetic storage, and was thus projected to catch up with magnetic storage by the year 2001. This suggests that by 2001, if nonvolatile DRAMs were developed, DRAMs might replace the magnetic hard disk as the permanent storage component of computers and change completely the computer system configuration as we know it today.

The results of the ICs supply model showed the most sensitivity to the changes in the number of silicon wafer defects per unit area. As expected, a decrease in the number of silicon wafer defects per unit area during the production of ICs increased the die yields and, ultimately, decreased the prices of DRAMs and CPUs. Since improved IC production learning and clean IC production environments can decrease the number of defects per unit area, efforts to make these improvements can lower the IC manufacturers' direct costs and, ultimately, increase the demand for workstations by allowing lower price-to-performance ratios.

5.1 Suggestions for Future Research

One possible extension to this dissertation would develop supply models of ICs to complement the super high-speed network communication. For instance, gallium arsenide, superconducting, or optical ICs supply models could be developed to study their effects on the performance and prices of future workstations and networks.

A second extension to this research could develop detailed supply and assembly models for liquid crystal displays and optical storage disks. Each of these components has several subcomponents and goes through several manufacturing stages before the final product is assembled. At each stage, there are several manufacturing procedures which must be performed, each with its own yield. Since these technologies are relatively new on the market, their production yield data are kept confidential by the manufacturers. Nevertheless, they will most likely replace the current market dominant CRT display and magnetic

storage technologies as soon as their manufacturers gain more experience in producing them; i.e., increase their yields, improve their response times, and reduce their costs.

A third extension to this dissertation might study the computer enhancements and the new applications that could result from replacing permanent magnetic storage with a permanent semiconductor one. DRAMs affect significantly the computer system's response time and performance. Moreover, they are simpler and more reliable semiconductor storage elements than their magnetic counterparts, and decreasing their cost per megabyte and improving their density could pave the way, for instance, to real-time computing. Real-time computing requires the computer job to finish within a time constraint or the job is canceled. If DRAMs were used as main memory and as permanent storage, the computer system's access to storage could be sped up by as much as 1 million times (20 milliseconds to 20 nanoseconds), making any storage intensive application look like a real-time application to the end user. Other new applications will be left for future research to study.

Since most of the computer technologies on the market carry with them uncertainties about their time to obsolescence and about the attributes of their successor technologies and the time of their introduction, further research could introduce uncertainty factors in the supply models and render their behavior stochastic.

A fifth and closely related project would study the life cycle of each introduced computer product by subjecting it to competitive forces from more innovative technologies emerging in the market place.

Further research launched from this dissertation could be performed at the company level, where sensitivity analyses might be done on the effects of new technologies introduced on the market. For instance, a sensitivity analysis might be designed to study the effects of RISC technology as the basis of all the processors and microcontrollers used in a workstation.

A linear workstation assembly process model was presented in this dissertation. The size of the model was not prohibitive from a computation time point of view (nevertheless, the process model can increase in size very quickly [46]). A seventh extension to this research would collect the data and include the following features in a future workstation assembly model:

- The major workstation manufacturers such as Sun, HP, IBM, DEC, Silicon Graphics, Sony, Toshiba, and Motorola.
- A list of regional markets corresponding to cities worldwide and associated transportation costs.
- Purchases of components and raw materials from around the world with corresponding transportation costs.
- Capacity expansion and depreciation factors of new plant and equipment, taking into consideration the expansion's investment constraints.
- A market share and life cycle analyses of old versus new workstation models, focusing especially on the time interval when the prices of the new models start to decrease.

These features might be particularly useful to workstation manufacturers to help them estimate where they stand in their production cycles of specific models and to help them define the workstation attributes they need to integrate to remain competitive in such a dynamic industry.

An eighth extension to this dissertation would study the effects of economies of scale on the supply of computer components and assembled workstations.

Since this dissertation is concerned with supply-side issues, a ninth extension to this research would study the demand side of computer components and assembled workstations.

5.2 Final Comments

“People and organizations resist change Shortage of funds may slow the growth progress, or the inability to increase the market enough to generate the necessary funds to develop technologies, or new applications do not get discovered In nature, exponential growth curves always top out. They turn into S-shaped curves ... the rapid, exponential expansion of the computer industry will slow down when it outstrips its resources, or whenever the market fails to supply exponentially increasing resources.” *W. Myers - 1991.*

This quotation by W. Myers [64] summarizes the state of the computer industry, the most dynamic industry of our time. By merging the concepts of cost and technical change, this dissertation captures the computer industry’s dynamic behavior and sets the stage for future research to be performed in this new field.

Appendix A

Samples of the Supply Models Outputs

The following appendix lists sample results of the CPU, DRAM, magnetic hard disk, and color CRT display supply models in Sections A.1 through A.5. The period of study ranges from 1980 to 1990, or 1985 to 1990, depending on the model. All the input parameters and all the assumptions made in Chapters 3 and 4 hold in all of the following simulation results.

A.1 CPU Supply Model Output

The organization of this section is as follows:

- Subsection A.1.1 presents the MIPS, speed, yield, and cost attributes of CISC CPUs for different die sizes. (Each table presents the results corresponding to one year of the study period.) No RISC CPUs results are reported because the following outputs represent the most pertinent samples and not the complete set of results that the model provides.
- Subsection A.1.2 reports the same results as those in Subsection A.1.1; however, each table lists the attributes corresponding to one of the die sizes chosen for the simulation.
- Subsection A.1.3 lists the die size, yield, and cost of fixed speed CPUs over the 1985-1990 period. Only the 1985-1990 results are reported because, again, the outputs represent samples of the model results.

A.1.1 CISC CPU MIPS, Speed and Cost Versus Die Size: 1980 - 1990

ICs - CISC CPUs: 1980

Die Area (cm ²)	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.774	0.05	1.56	1.995	0.42	7.55
0.2601	0.513	0.20	6.50	1.995	2.89	10.94
0.5776	0.304	0.44	14.44	1.995	11.99	22.36
1.0404	0.170	0.78	26.01	1.995	43.66	60.06
1.3064	0.130	0.98	32.66	1.995	76.77	150.85
1.6129	0.099	1.21	40.32	1.995	134.12	216.91
2.3104	0.060	1.74	57.76	1.995	379.43	495.98
3.1684	0.037	2.39	79.21	1.995	1063.36	1266.14
4.1209	0.024	3.10	103.02	1.995	2852.77	3268.68

ICs - CISC CPUs: 1981

Die Area (cm ²)	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.775	0.06	1.62	1.758	0.39	7.59
0.2601	0.514	0.26	6.75	1.758	2.68	10.80
0.5776	0.305	0.58	15.00	1.758	11.03	21.45
1.0404	0.171	1.05	27.02	1.758	39.65	55.88
1.3064	0.131	1.32	33.92	1.758	69.24	142.82
1.6129	0.100	1.63	41.88	1.758	119.96	201.62
2.3104	0.060	2.33	60.00	1.758	332.39	444.44
3.1684	0.037	3.19	82.28	1.758	901.88	1087.89
4.1209	0.024	4.15	107.01	1.758	2297.18	2653.13

ICs - CISC CPUs: 1982

Die Area (cm ²)	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.774	0.09	1.71	1.549	0.37	7.69
0.2601	0.508	0.35	7.10	1.549	2.52	10.83
0.5776	0.295	0.79	15.76	1.549	10.52	21.32
1.0404	0.160	1.42	28.40	1.549	38.83	55.98
1.3064	0.120	1.78	35.66	1.549	68.77	143.82
1.6129	0.089	2.20	44.02	1.549	120.95	204.96
2.3104	0.051	3.15	63.06	1.549	344.43	462.49
3.1684	0.030	4.32	86.48	1.549	954.40	1155.52
4.1209	0.019	5.62	112.47	1.549	2443.87	2832.90

ICs - CISC CPUs: 1983

Die Area (cm ²)	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.774	0.11	1.77	1.365	0.34	7.73
0.2601	0.510	0.47	7.37	1.365	2.33	10.73
0.5776	0.296	1.05	16.38	1.365	9.65	20.52
1.0404	0.161	1.90	29.50	1.365	35.23	52.30
1.3064	0.121	2.39	37.04	1.365	62.05	136.76
1.6129	0.090	2.95	45.73	1.365	108.42	191.60
2.3104	0.052	4.22	65.50	1.365	304.08	418.60
3.1684	0.030	5.79	89.82	1.365	824.69	1012.98
4.1209	0.019	7.53	116.83	1.365	2049.31	2397.00

ICs - CISC CPUs: 1984

Die Area (cm ²)	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.774	0.15	1.86	1.202	0.32	7.84
0.2601	0.505	0.64	7.74	1.202	2.18	10.78
0.5776	0.288	1.43	17.19	1.202	9.13	20.42
1.0404	0.151	2.57	30.96	1.202	34.17	52.24
1.3064	0.112	3.23	38.88	1.202	61.00	137.30
1.6129	0.082	3.98	48.00	1.202	108.16	193.90
2.3104	0.045	5.71	68.76	1.202	312.27	433.33
3.1684	0.025	7.82	94.29	1.202	869.80	1074.85
4.1209	0.015	10.18	122.64	1.202	2201.29	2588.16

ICs - CISC CPUs: 1985

Die Area (cm ²)	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.774	0.21	1.93	1.059	0.30	7.90
0.2601	0.507	0.86	8.04	1.059	2.01	10.72
0.5776	0.290	1.91	17.85	1.059	8.35	19.77
1.0404	0.153	3.44	32.16	1.059	30.97	49.07
1.3064	0.112	4.32	40.38	1.059	55.03	131.18
1.6129	0.082	5.33	49.86	1.059	97.07	182.26
2.3104	0.046	7.63	71.42	1.059	276.99	395.32
3.1684	0.026	10.47	97.94	1.059	759.76	954.67
4.1209	0.015	13.62	127.39	1.059	1885.59	2240.81

ICs - CISC CPUs: 1986

Die Area (cm ²)	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.773	0.28	2.03	0.933	0.27	8.02
0.2601	0.503	1.16	8.43	0.933	1.86	10.81
0.5776	0.283	2.58	18.72	0.933	7.86	19.73
1.0404	0.145	4.64	33.73	0.933	29.79	49.00
1.3064	0.105	5.83	42.35	0.933	53.61	131.52
1.6129	0.076	7.20	52.29	0.933	95.93	183.96
2.3104	0.040	10.31	74.90	0.933	281.84	407.43
3.1684	0.022	14.14	102.71	0.933	795.96	1009.56
4.1209	0.012	18.40	133.59	0.933	2023.88	2423.53

ICs - CISC CPUs: 1987

Die Area (cm ²)	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.773	0.38	2.12	0.822	0.25	8.15
0.2601	0.500	1.57	8.83	0.822	1.73	10.91
0.5776	0.277	3.48	19.62	0.822	7.37	19.71
1.0404	0.138	6.27	35.34	0.822	28.47	48.86
1.3064	0.099	7.87	44.37	0.822	51.85	131.62
1.6129	0.070	9.72	54.78	0.822	94.02	185.08
2.3104	0.036	13.92	78.48	0.822	284.03	417.49
3.1684	0.019	19.09	107.62	0.822	825.61	1059.91
4.1209	0.010	24.83	139.97	0.822	2152.75	2602.72

ICs - CISC CPUs: 1988

Die Area (cm ²)	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.773	0.50	2.21	0.724	0.23	8.24
0.2601	0.501	2.10	9.18	0.724	1.59	10.91
0.5776	0.278	4.66	20.38	0.724	6.72	19.25
1.0404	0.139	8.39	36.71	0.724	25.76	46.35
1.3064	0.100	10.53	46.09	0.724	46.74	126.63
1.6129	0.071	13.01	56.91	0.724	84.41	175.35
2.3104	0.036	18.63	81.51	0.724	252.83	384.59
3.1684	0.019	25.55	111.79	0.724	727.37	954.06
4.1209	0.010	33.23	145.39	0.724	1874.18	2299.05

ICs - CISC CPUs: 1989

Die Area (cm ²)	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.773	0.68	2.31	0.638	0.22	8.38
0.2601	0.498	2.83	9.61	0.638	1.47	11.05
0.5776	0.273	6.28	21.34	0.638	6.26	19.33
1.0404	0.134	11.32	38.43	0.638	24.47	46.38
1.3064	0.094	14.21	48.26	0.638	44.89	126.87
1.6129	0.066	17.54	59.58	0.638	82.11	176.46
2.3104	0.033	25.13	85.35	0.638	252.74	393.40
3.1684	0.016	34.46	117.05	0.638	748.75	998.95
4.1209	0.009	44.82	152.24	0.638	1982.26	2464.62

ICs - CISC CPUs: 1990						
Die Area (cm ²)	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.773	0.92	2.42	0.562	0.20	8.54
0.2601	0.496	3.81	10.06	0.562	1.35	11.21
0.5776	0.269	8.47	22.33	0.562	5.83	19.45
1.0404	0.129	15.26	40.22	0.562	23.13	46.43
1.3064	0.090	19.16	50.50	0.562	42.88	127.08
1.6129	0.062	23.65	62.35	0.562	79.36	177.36
2.3104	0.030	33.88	89.32	0.562	250.68	401.07
3.1684	0.014	46.46	122.49	0.562	764.09	1040.54
4.1209	0.007	60.42	159.31	0.562	2078.07	2625.95

A.1.2 CPU MIPS, Speed and Cost Versus Year: 0.0625 cm² - 4.1209 cm²

ICs - CISC CPUs: Die Area = 0.0625 cm²

Year	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
1980	0.774	0.05	1.56	1.995	0.42	7.55
1981	0.775	0.06	1.62	1.758	0.39	7.59
1982	0.774	0.09	1.71	1.549	0.37	7.69
1983	0.774	0.11	1.77	1.365	0.34	7.73
1984	0.774	0.15	1.86	1.202	0.32	7.84
1985	0.774	0.21	1.93	1.059	0.30	7.90
1986	0.773	0.28	2.03	0.933	0.27	8.02
1987	0.773	0.38	2.12	0.822	0.25	8.15
1988	0.773	0.50	2.21	0.724	0.23	8.24
1989	0.773	0.68	2.31	0.638	0.22	8.38
1990	0.773	0.92	2.42	0.562	0.20	8.54

ICs - CISC CPUs: Die Area = 0.2601 cm²

Year	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
1980	0.513	0.20	6.50	1.995	2.89	10.94
1981	0.514	0.26	6.75	1.758	2.68	10.80
1982	0.508	0.35	7.10	1.549	2.52	10.83
1983	0.510	0.47	7.37	1.365	2.33	10.73
1984	0.505	0.64	7.74	1.202	2.18	10.78
1985	0.507	0.86	8.04	1.059	2.01	10.72
1986	0.503	1.16	8.43	0.933	1.86	10.81
1987	0.500	1.57	8.83	0.822	1.73	10.91
1988	0.501	2.10	9.18	0.724	1.59	10.91
1989	0.498	2.83	9.61	0.638	1.47	11.05
1990	0.496	3.81	10.06	0.562	1.35	11.21

ICs - CISC CPUs: Die Area = 0.5776 cm²

Year	Die	MIPS	Speed	Feature	Manuf.	Total
	yield		(MHz)	Size	Cost	Cost
	(%)			(micron)	(\$)	(\$)
1980	0.304	0.44	14.44	1.995	11.99	22.36
1981	0.305	0.58	15.00	1.758	11.03	21.45
1982	0.295	0.79	15.76	1.549	10.52	21.32
1983	0.296	1.05	16.38	1.365	9.65	20.52
1984	0.288	1.43	17.19	1.202	9.13	20.42
1985	0.290	1.91	17.85	1.059	8.35	19.77
1986	0.283	2.58	18.72	0.933	7.86	19.73
1987	0.277	3.48	19.62	0.822	7.37	19.71
1988	0.278	4.66	20.38	0.724	6.72	19.25
1989	0.273	6.28	21.34	0.638	6.26	19.33
1990	0.269	8.47	22.33	0.562	5.83	19.45

ICs - CISC CPUs: Die Area = 1.0404 cm²

Year	Die	MIPS	Speed	Feature	Manuf.	Total
	yield		(MHz)	Size	Cost	Cost
	(%)			(micron)	(\$)	(\$)
1980	0.170	0.78	26.01	1.995	43.66	60.06
1981	0.171	1.05	27.02	1.758	39.65	55.88
1982	0.160	1.42	28.40	1.549	38.83	55.98
1983	0.161	1.90	29.50	1.365	35.23	52.30
1984	0.151	2.57	30.96	1.202	34.17	52.24
1985	0.153	3.44	32.16	1.059	30.97	49.07
1986	0.145	4.64	33.73	0.933	29.79	49.00
1987	0.138	6.27	35.34	0.822	28.47	48.86
1988	0.139	8.39	36.71	0.724	25.76	46.35
1989	0.134	11.32	38.43	0.638	24.47	46.38
1990	0.129	15.26	40.22	0.562	23.13	46.43

ICs - CISC CPUs: Die Area = 1.3064 cm²

Year	Die	MIPS	Speed	Feature	Manuf.	Total
	yield		(MHz)	Size	Cost	Cost
	(%)			(micron)	(\$)	(\$)
1980	0.130	0.98	32.66	1.995	76.77	150.85
1981	0.131	1.32	33.92	1.758	69.24	142.82
1982	0.120	1.78	35.66	1.549	68.77	143.82
1983	0.121	2.39	37.04	1.365	62.05	136.76
1984	0.112	3.23	38.88	1.202	61.00	137.30
1985	0.112	4.32	40.38	1.059	55.03	131.18
1986	0.105	5.83	42.35	0.933	53.61	131.52
1987	0.099	7.87	44.37	0.822	51.85	131.62
1988	0.100	10.53	46.09	0.724	46.74	126.63
1989	0.094	14.21	48.26	0.638	44.89	126.87
1990	0.090	19.16	50.50	0.562	42.88	127.08

ICs - CISC CPUs: Die Area = 1.6129 cm²

Year	Die	MIPS	Speed	Feature	Manuf.	Total
	yield		(MHz)	Size	Cost	Cost
	(%)			(micron)	(\$)	(\$)
1980	0.099	1.21	40.32	1.995	134.12	216.91
1981	0.100	1.63	41.88	1.758	119.96	201.62
1982	0.089	2.20	44.02	1.549	120.95	204.96
1983	0.090	2.95	45.73	1.365	108.42	191.60
1984	0.082	3.98	48.00	1.202	108.16	193.90
1985	0.082	5.33	49.86	1.059	97.07	182.26
1986	0.076	7.20	52.29	0.933	95.93	183.96
1987	0.070	9.72	54.78	0.822	94.02	185.08
1988	0.071	13.01	56.91	0.724	84.41	175.35
1989	0.066	17.54	59.58	0.638	82.11	176.46
1990	0.062	23.65	62.35	0.562	79.36	177.36

ICs - CISC CPUs: Die Area = 2.3104 cm²

Year	Die	MIPS	Speed	Feature	Manuf.	Total
	yield		(MHz)	Size	Cost	Cost
	(%)			(micron)	(\\$)	(\\$)
1980	0.060	1.74	57.76	1.995	379.43	495.98
1981	0.060	2.33	60.00	1.758	332.39	444.44
1982	0.051	3.15	63.06	1.549	344.43	462.49
1983	0.052	4.22	65.50	1.365	304.08	418.60
1984	0.045	5.71	68.76	1.202	312.27	433.33
1985	0.046	7.63	71.42	1.059	276.99	395.32
1986	0.040	10.31	74.90	0.933	281.84	407.43
1987	0.036	13.92	78.48	0.822	284.03	417.49
1988	0.036	18.63	81.51	0.724	252.83	384.59
1989	0.033	25.13	85.35	0.638	252.74	393.40
1990	0.030	33.88	89.32	0.562	250.68	401.07

ICs - CISC CPUs: Die Area = 3.1684 cm²

Year	Die	MIPS	Speed	Feature	Manuf.	Total
	yield		(MHz)	Size	Cost	Cost
	(%)			(micron)	(\\$)	(\\$)
1980	0.037	2.39	79.21	1.995	1063.36	1266.14
1981	0.037	3.19	82.28	1.758	901.88	1087.89
1982	0.030	4.32	86.48	1.549	954.40	1155.52
1983	0.030	5.79	89.82	1.365	824.69	1012.98
1984	0.025	7.82	94.29	1.202	869.80	1074.85
1985	0.026	10.47	97.94	1.059	759.76	954.67
1986	0.022	14.14	102.71	0.933	795.96	1009.56
1987	0.019	19.09	107.62	0.822	825.61	1059.91
1988	0.019	25.55	111.79	0.724	727.37	954.06
1989	0.016	34.46	117.05	0.638	748.75	998.95
1990	0.014	46.46	122.49	0.562	764.09	1040.54

ICs - CISC CPUs: Die Area = 4.1209 cm²

Year	Die yield (%)	MIPS	Speed (MHz)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
1980	0.024	3.10	103.02	1.995	2852.77	3268.68
1981	0.024	4.15	107.01	1.758	2297.18	2653.13
1982	0.019	5.62	112.47	1.549	2443.87	2832.90
1983	0.019	7.53	116.83	1.365	2049.31	2397.00
1984	0.015	10.18	122.64	1.202	2201.29	2588.16
1985	0.015	13.62	127.39	1.059	1885.59	2240.81
1986	0.012	18.40	133.59	0.933	2023.88	2423.53
1987	0.010	24.83	139.97	0.822	2152.75	2602.72
1988	0.010	33.23	145.39	0.724	1874.18	2299.05
1989	0.009	44.82	152.24	0.638	1982.26	2464.62
1990	0.007	60.42	159.31	0.562	2078.07	2625.95

A.1.3 Fixed Speed CPU Cost and Die Area: 1985 - 1990

Fixed Speed CISC CPU: 20 MHz

Year	Die yield (%)	Die Area (cm ²)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)	Cost/MHz (\$)
1985	0.260	0.65	1.059	10.55	22.74	1.14
1986	0.265	0.62	0.933	9.01	21.35	1.07
1987	0.272	0.59	0.822	7.67	20.15	1.01
1988	0.283	0.57	0.724	6.46	18.86	0.94
1989	0.291	0.54	0.638	5.48	18.06	0.90
1990	0.300	0.52	0.562	4.64	17.40	0.87

Fixed Speed CISC CPU: 40 MHz

Year	Die yield (%)	Die Area (cm ²)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)	Cost/MHz (\$)
1985	0.114	1.29	1.059	53.68	129.51	3.24
1986	0.114	1.23	0.933	46.09	122.04	3.05
1987	0.116	1.18	0.822	39.23	115.27	2.88
1988	0.123	1.13	0.724	32.09	107.15	2.68
1989	0.126	1.08	0.638	27.12	50.13	1.25
1990	0.130	1.03	0.562	22.80	45.94	1.15

Fixed Speed CISC CPU: 60 MHz

Year	Die yield (%)	Die Area (cm ²)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)	Cost/MHz (\$)
1985	0.061	1.94	1.059	164.48	262.77	4.38
1986	0.060	1.85	0.933	143.10	241.39	4.02
1987	0.060	1.77	0.822	123.03	221.21	3.69
1988	0.064	1.70	0.724	98.52	193.29	3.22
1989	0.065	1.62	0.638	83.82	178.70	2.98
1990	0.066	1.55	0.562	70.70	165.59	2.76

A.2 DRAM Supply Model Output

The organization of this section is as follows:

- Subsection A.2.1 presents the DRAM capacity, yield, and cost attributes for different die sizes. (Each table presents the results corresponding to one year of the study period.)
- Subsection A.2.2 reports the same results as those in Subsection A.2.1; however, each table lists the attributes corresponding to one of the die sizes chosen for the simulation.
- Subsection A.2.3 lists the die size, yield, and cost of fixed capacity DRAMs over the 1985-1990 period. Only the 1985-1990 results are reported because the outputs represent samples of the model results and not the complete set.

A.2.1 DRAM Capacity and Cost Versus Die Size: 1980 - 1990

ICs - DRAMs: 1980

Die Area (cm ²)	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.774	0.002	1.995	0.42	7.55
0.2601	0.513	0.006	1.995	2.89	10.94
0.5776	0.304	0.014	1.995	11.99	22.36
1.0404	0.170	0.025	1.995	43.66	60.06
1.3064	0.130	0.033	1.995	76.77	150.85
1.6129	0.099	0.040	1.995	134.12	216.91
2.3104	0.060	0.058	1.995	379.43	495.98
3.1684	0.037	0.079	1.995	1063.36	1266.14
4.1209	0.024	0.103	1.995	2852.77	3268.68

ICs - DRAMs: 1981

Die Area (cm ²)	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.775	0.002	1.758	0.39	7.59
0.2601	0.514	0.008	1.758	2.68	10.80
0.5776	0.305	0.019	1.758	11.03	21.45
1.0404	0.171	0.032	1.758	39.65	55.88
1.3064	0.131	0.042	1.758	69.24	142.82
1.6129	0.100	0.052	1.758	119.96	201.62
2.3104	0.060	0.074	1.758	332.39	444.44
3.1684	0.037	0.102	1.758	901.88	1087.89
4.1209	0.024	0.133	1.758	2297.18	2653.13

ICs - DRAMs: 1982

Die Area (cm ²)	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.774	0.003	1.549	0.37	7.69
0.2601	0.508	0.014	1.549	2.52	10.83
0.5776	0.295	0.030	1.549	10.52	21.32
1.0404	0.160	0.053	1.549	38.83	55.98
1.3064	0.120	0.069	1.549	68.77	143.82
1.6129	0.089	0.085	1.549	120.95	204.96
2.3104	0.051	0.121	1.549	344.43	462.49
3.1684	0.030	0.166	1.549	954.40	1155.52
4.1209	0.019	0.216	1.549	2443.87	2832.90

ICs - DRAMs: 1983

Die Area (cm ²)	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.774	0.004	1.365	0.34	7.73
0.2601	0.510	0.018	1.365	2.33	10.73
0.5776	0.296	0.039	1.365	9.65	20.52
1.0404	0.161	0.068	1.365	35.23	52.30
1.3064	0.121	0.088	1.365	62.05	136.76
1.6129	0.090	0.109	1.365	108.42	191.60
2.3104	0.052	0.156	1.365	304.08	418.60
3.1684	0.030	0.215	1.365	824.69	1012.98
4.1209	0.019	0.279	1.365	2049.31	2397.00

ICs - DRAMs: 1984

Die Area (cm ²)	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.774	0.007	1.202	0.32	7.84
0.2601	0.505	0.028	1.202	2.18	10.78
0.5776	0.288	0.062	1.202	9.13	20.42
1.0404	0.151	0.107	1.202	34.17	52.24
1.3064	0.112	0.140	1.202	61.00	137.30
1.6129	0.082	0.173	1.202	108.16	193.90
2.3104	0.045	0.249	1.202	312.27	433.33
3.1684	0.025	0.341	1.202	869.80	1074.85
4.1209	0.015	0.443	1.202	2201.29	2588.16

ICs - DRAMs: 1985

Die Area (cm ²)	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.774	0.009	1.059	0.30	7.90
0.2601	0.507	0.036	1.059	2.01	10.72
0.5776	0.290	0.080	1.059	8.35	19.77
1.0404	0.153	0.139	1.059	30.97	49.07
1.3064	0.112	0.181	1.059	55.03	131.18
1.6129	0.082	0.223	1.059	97.07	182.26
2.3104	0.046	0.320	1.059	276.99	395.32
3.1684	0.026	0.440	1.059	759.76	954.67
4.1209	0.015	0.571	1.059	1885.59	2240.81

ICs - DRAMs: 1986

Die Area (cm ²)	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.773	0.013	0.933	0.27	8.02
0.2601	0.503	0.056	0.933	1.86	10.81
0.5776	0.283	0.125	0.933	7.86	19.73
1.0404	0.145	0.216	0.933	29.79	49.00
1.3064	0.105	0.282	0.933	53.61	131.52
1.6129	0.076	0.348	0.933	95.93	183.96
2.3104	0.040	0.450	0.933	281.84	407.43
3.1684	0.022	0.684	0.933	795.96	1009.56
4.1209	0.012	0.890	0.933	2023.88	2423.53

ICs - DRAMs: 1987

Die Area (cm ²)	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.773	0.021	0.822	0.25	8.15
0.2601	0.500	0.086	0.822	1.73	10.91
0.5776	0.277	0.191	0.822	7.37	19.71
1.0404	0.138	0.331	0.822	28.47	48.86
1.3064	0.099	0.433	0.822	51.85	131.62
1.6129	0.070	0.534	0.822	94.02	185.08
2.3104	0.036	0.765	0.822	284.03	417.49
3.1684	0.019	1.049	0.822	825.61	1059.91
4.1209	0.010	1.365	0.822	2152.75	2602.72

ICs - DRAMs: 1988

Die Area (cm ²)	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.773	0.027	0.724	0.23	8.24
0.2601	0.501	0.111	0.724	1.59	10.91
0.5776	0.278	0.250	0.724	6.72	19.25
1.0404	0.139	0.427	0.724	25.76	46.35
1.3064	0.100	0.560	0.724	46.74	126.63
1.6129	0.071	0.690	0.724	84.41	175.35
2.3104	0.036	0.990	0.724	252.83	384.59
3.1684	0.019	1.352	0.724	727.37	954.06
4.1209	0.010	1.760	0.724	1874.18	2299.05

ICs - DRAMs: 1989

Die Area (cm ²)	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.773	0.040	0.638	0.22	8.38
0.2601	0.498	0.170	0.638	1.47	11.05
0.5776	0.273	0.373	0.638	6.26	19.33
1.0404	0.134	0.670	0.638	24.47	46.38
1.3064	0.094	0.843	0.638	44.89	126.87
1.6129	0.066	1.041	0.638	82.11	176.46
2.3104	0.033	1.491	0.638	252.74	393.40
3.1684	0.016	2.044	0.638	748.75	998.95
4.1209	0.009	2.660	0.638	1982.26	2464.62

ICs - DRAMs: 1990

Die Area (cm ²)	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
0.0625	0.773	0.060	0.562	0.20	8.54
0.2601	0.496	0.251	0.562	1.35	11.21
0.5776	0.269	0.560	0.562	5.83	19.45
1.0404	0.129	1.000	0.562	23.13	46.43
1.3064	0.090	1.260	0.562	42.88	127.08
1.6129	0.062	1.555	0.562	79.36	177.36
2.3104	0.030	2.230	0.562	250.68	401.07
3.1684	0.014	3.054	0.562	764.09	1040.54
4.1209	0.007	3.972	0.562	2078.07	2625.95

A.2.2 DRAM Capacity and Cost Versus Year: 0.0625 cm² - 4.1209 cm²

ICs - DRAMs: Die Area = 0.0625 cm²

Year	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
1980	0.774	0.002	1.995	0.42	7.55
1981	0.775	0.002	1.758	0.39	7.59
1982	0.774	0.003	1.549	0.37	7.69
1983	0.774	0.004	1.365	0.34	7.73
1984	0.774	0.007	1.202	0.32	7.84
1985	0.774	0.009	1.059	0.30	7.90
1986	0.773	0.013	0.933	0.27	8.02
1987	0.773	0.021	0.822	0.25	8.15
1988	0.773	0.027	0.724	0.23	8.24
1989	0.773	0.040	0.638	0.22	8.38
1990	0.773	0.060	0.562	0.20	8.54

ICs - DRAMs: Die Area = 0.2601 cm²

Year	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
1980	0.513	0.006	1.995	2.89	10.94
1981	0.514	0.008	1.758	2.68	10.80
1982	0.508	0.014	1.549	2.52	10.83
1983	0.510	0.018	1.365	2.33	10.73
1984	0.505	0.028	1.202	2.18	10.78
1985	0.507	0.036	1.059	2.01	10.72
1986	0.503	0.056	0.933	1.86	10.81
1987	0.500	0.086	0.822	1.73	10.91
1988	0.501	0.111	0.724	1.59	10.91
1989	0.498	0.170	0.638	1.47	11.05
1990	0.496	0.251	0.562	1.35	11.21

ICs - DRAMs: Die Area = 0.5776 cm²

Year	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
1980	0.304	0.014	1.995	11.99	22.36
1981	0.305	0.019	1.758	11.03	21.45
1982	0.295	0.030	1.549	10.52	21.32
1983	0.296	0.039	1.365	9.65	20.52
1984	0.288	0.062	1.202	9.13	20.42
1985	0.290	0.080	1.059	8.35	19.77
1986	0.283	0.125	0.933	7.86	19.73
1987	0.277	0.191	0.822	7.37	19.71
1988	0.278	0.250	0.724	6.72	19.25
1989	0.273	0.373	0.638	6.26	19.33
1990	0.269	0.560	0.562	5.83	19.45

ICs - DRAMs: Die Area = 1.0404 cm²

Year	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
1980	0.170	0.025	1.995	43.66	60.06
1981	0.171	0.032	1.758	39.65	55.88
1982	0.160	0.053	1.549	38.83	55.98
1983	0.161	0.068	1.365	35.23	52.30
1984	0.151	0.107	1.202	34.17	52.24
1985	0.153	0.139	1.059	30.97	49.07
1986	0.145	0.216	0.933	29.79	49.00
1987	0.138	0.331	0.822	28.47	48.86
1988	0.139	0.427	0.724	25.76	46.35
1989	0.134	0.670	0.638	24.47	46.38
1990	0.129	1.000	0.562	23.13	46.43

ICs - DRAMs: Die Area = 1.3064 cm²

Year	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
1980	0.130	0.033	1.995	76.77	150.85
1981	0.131	0.042	1.758	69.24	142.82
1982	0.120	0.069	1.549	68.77	143.82
1983	0.121	0.088	1.365	62.05	136.76
1984	0.112	0.140	1.202	61.00	137.30
1985	0.112	0.181	1.059	55.03	131.18
1986	0.105	0.282	0.933	53.61	131.52
1987	0.099	0.433	0.822	51.85	131.62
1988	0.100	0.560	0.724	46.74	126.63
1989	0.094	0.843	0.638	44.89	126.87
1990	0.090	1.260	0.562	42.88	127.08

ICs - DRAMs: Die Area = 1.6129 cm²

Year	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
1980	0.099	0.040	1.995	134.12	216.91
1981	0.100	0.052	1.758	119.96	201.62
1982	0.089	0.085	1.549	120.95	204.96
1983	0.090	0.109	1.365	108.42	191.60
1984	0.082	0.173	1.202	108.16	193.90
1985	0.082	0.223	1.059	97.07	182.26
1986	0.076	0.348	0.933	95.93	183.96
1987	0.070	0.534	0.822	94.02	185.08
1988	0.071	0.690	0.724	84.41	175.35
1989	0.066	1.041	0.638	82.11	176.46
1990	0.062	1.555	0.562	79.36	177.36

ICs - DRAMs: Die Area = 2.3104 cm²

Year	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
1980	0.060	0.058	1.995	379.43	495.98
1981	0.060	0.074	1.758	332.39	444.44
1982	0.051	0.121	1.549	344.43	462.49
1983	0.052	0.156	1.365	304.08	418.60
1984	0.045	0.249	1.202	312.27	433.33
1985	0.046	0.320	1.059	276.99	395.32
1986	0.040	0.450	0.933	281.84	407.43
1987	0.036	0.765	0.822	284.03	417.49
1988	0.036	0.990	0.724	252.83	384.59
1989	0.033	1.491	0.638	252.74	393.40
1990	0.030	2.230	0.562	250.68	401.07

ICs - DRAMs: Die Area = 3.1684 cm²

Year	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
1980	0.037	0.079	1.995	1063.36	1266.14
1981	0.037	0.102	1.758	901.88	1087.89
1982	0.030	0.166	1.549	954.40	1155.52
1983	0.030	0.215	1.365	824.69	1012.98
1984	0.025	0.341	1.202	869.80	1074.85
1985	0.026	0.440	1.059	759.76	954.67
1986	0.022	0.684	0.933	795.96	1009.56
1987	0.019	1.049	0.822	825.61	1059.91
1988	0.019	1.352	0.724	727.37	954.06
1989	0.016	2.044	0.638	748.75	998.95
1990	0.014	3.054	0.562	764.09	1040.54

ICs - DRAMs: Die Area = 4.1209 cm ²					
Year	Die yield (%)	DRAM (MB)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)
1980	0.024	0.103	1.995	2852.77	3268.68
1981	0.024	0.133	1.758	2297.18	2653.13
1982	0.019	0.216	1.549	2443.87	2832.90
1983	0.019	0.279	1.365	2049.31	2397.00
1984	0.015	0.443	1.202	2201.29	2588.16
1985	0.015	0.571	1.059	1885.59	2240.81
1986	0.012	0.890	0.933	2023.88	2423.53
1987	0.010	1.365	0.822	2152.75	2602.72
1988	0.010	1.760	0.724	1874.18	2299.05
1989	0.009	2.660	0.638	1982.26	2464.62
1990	0.007	3.972	0.562	2078.07	2625.95

A.2.3 Fixed Capacity DRAM Cost and Die Area: 1985 -1990

Fixed Capacity DRAM: 0.1 MB						
Year	Die yield (%)	Die Area (cm ²)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)	Cost/MB (\$)
1985	0.233	0.72	1.059	13.32	26.41	264.14
1986	0.343	0.46	0.933	5.08	15.72	157.22
1987	0.459	0.30	0.822	2.21	11.72	117.17
1988	0.528	0.23	0.724	1.35	10.47	104.70
1989	0.625	0.16	0.638	0.68	9.45	94.51
1990	0.701	0.10	0.562	0.37	8.98	89.85

Fixed Capacity DRAM: 0.2 MB

Year	Die yield (%)	Die Area (cm ²)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)	Cost/MB (\$)
1985	0.097	1.44	1.059	71.62	151.46	757.32
1986	0.168	0.93	0.933	22.39	39.37	196.87
1987	0.265	0.60	0.822	8.08	20.76	103.80
1988	0.336	0.47	0.724	4.43	15.69	78.45
1989	0.450	0.31	0.638	1.96	11.98	59.89
1990	0.555	0.21	0.562	0.96	10.36	51.78

Fixed Capacity DRAM: 0.3 MB

Year	Die yield (%)	Die Area (cm ²)	Feature Size (micron)	Manuf. Cost (\$)	Total Cost (\$)	Cost/MB (\$)
1985	0.051	2.16	1.059	227.24	336.91	1123.02
1986	0.096	1.39	0.933	63.21	143.53	478.43
1987	0.167	0.91	0.822	20.19	37.73	125.78
1988	0.227	0.70	0.724	10.21	24.48	81.61
1989	0.334	0.47	0.638	4.06	15.69	52.30
1990	0.446	0.31	0.562	1.82	12.16	40.54

A.3 Magnetic Hard Disk Supply Model Output

The organization of this section is as follows:

- Subsection A.3.1 presents the magnetic hard disk cost per megabyte results.
- Subsection A.3.2 provides the behavior, over time, of the magnetic hard disk areal density.
- Subsection A.3.3 lists the areal capacity, the volumetric capacity, the data rate, and the cost of several magnetic hard disk configurations. Each disk configuration includes a certain disk diameter and a storage system height.

A.3.1 Magnetic Hard Disk Cost/MB

Magnetic Hard Disk	
Year	Cost/MB (\$)
1980	18.74
1981	15.40
1982	12.51
1983	10.28
1984	8.36
1985	6.87
1986	5.59
1987	4.55
1988	3.74
1989	3.05
1990	2.49

A.3.2 Magnetic Hard Disk Areal Density

Magnetic Hard Disk		
Year	Megabit/cm ² (Mb/cm ²)	Megabyte/cm ² (MB/cm ²)
1980	1.27	0.16
1981	1.65	0.21
1982	2.15	0.27
1983	2.81	0.35
1984	3.66	0.46
1985	4.77	0.60
1986	6.21	0.78
1987	8.10	1.01
1988	10.55	1.32
1989	13.75	1.72
1990	17.92	2.24

A.3.3 Magnetic Hard Disk Parameters for Different Diameter and Height Specifications

HD: D = 1.8 in, H = 1 in

Year	Areal Capacity (MB)	Volumetric Capacity (MB)	Data Rate (MB/sec)	Total Hard Disk Cost (\$)
1980	1.21	2.43	0.18	45.505
1981	1.57	3.15	0.21	48.458
1982	2.05	4.09	0.24	51.185
1983	2.66	5.31	0.28	54.580
1984	3.45	6.90	0.32	57.641
1985	4.48	8.95	0.37	61.477
1986	5.81	11.62	0.43	64.909
1987	7.54	15.09	0.49	68.678
1988	9.79	39.16	1.13	146.465
1989	12.70	50.82	1.30	154.928
1990	16.49	65.95	1.49	163.992

HD: D = 1.8 in, H = 2 in

Year	Areal Capacity (MB)	Volumetric Capacity (MB)	Data Rate (MB/sec)	Total Hard Disk Cost (\$)
1980	1.21	4.86	0.36	91.009
1981	1.57	6.29	0.42	96.916
1982	2.05	8.18	0.48	102.369
1983	2.66	10.62	0.56	109.159
1984	3.45	13.80	0.64	115.281
1985	4.48	26.86	1.11	184.431
1986	5.81	34.85	1.28	194.728
1987	7.54	45.26	1.47	206.033
1988	9.79	78.32	2.26	292.930
1989	12.70	101.64	2.59	309.856
1990	16.49	164.89	3.72	409.980

HD: D = 3.5 in, H = 1 in

Year	Areal Capacity (MB)	Volumetric Capacity (MB)	Data Rate (MB/sec)	Total Hard Disk Cost (\$)
1980	4.57	9.15	0.35	171.432
1981	5.94	11.89	0.41	183.119
1982	7.72	15.44	0.47	193.106
1983	10.02	20.04	0.54	205.976
1984	13.02	26.03	0.63	217.521
1985	16.90	33.81	0.72	232.126
1986	21.95	43.90	0.83	245.319
1987	28.50	57.01	0.95	259.473
1988	37.00	148.00	2.19	553.520
1989	48.03	192.12	2.52	585.709
1990	62.32	249.28	2.89	619.827

HD: D = 3.5 in, H = 2 in

Year	Areal Capacity (MB)	Volumetric Capacity (MB)	Data Rate (MB/sec)	Total Hard Disk Cost (\$)
1980	4.57	18.29	0.70	342.864
1981	5.94	23.78	0.81	366.238
1982	7.72	30.88	0.94	386.212
1983	10.02	40.08	1.08	411.952
1984	13.02	52.06	1.25	435.043
1985	16.90	101.43	2.16	696.378
1986	21.95	131.70	2.49	735.957
1987	28.50	171.02	2.86	778.418
1988	37.00	296.00	4.39	1107.040
1989	48.03	384.25	5.04	1171.417
1990	62.32	623.21	7.24	1549.567

HD: D = 5.25 in, H = 1 in

Year	Areal Capacity (MB)	Volumetric Capacity (MB)	Data Rate (MB/sec)	Total Hard Disk Cost (\$)
1980	10.28	20.57	0.53	385.525
1981	13.36	26.73	0.61	411.654
1982	17.35	34.70	0.70	433.993
1983	22.54	45.08	0.81	463.294
1984	29.27	58.54	0.94	489.147
1985	38.03	76.05	1.08	522.156
1986	49.37	98.74	1.24	551.736
1987	64.11	128.22	1.43	583.602
1988	83.22	332.89	3.29	1245.031
1989	108.03	432.11	3.78	1317.313
1990	140.20	560.82	4.34	1394.449

HD: D = 5.25 in, H = 2 in

Year	Areal Capacity (MB)	Volumetric Capacity (MB)	Data Rate (MB/sec)	Total Hard Disk Cost (\$)
1980	10.28	41.13	1.05	771.049
1981	13.36	53.45	1.22	823.308
1982	17.35	69.40	1.41	867.985
1983	22.54	90.16	1.63	926.588
1984	29.27	117.08	1.88	978.293
1985	38.03	228.15	3.24	1566.468
1986	49.37	296.21	3.73	1655.208
1987	64.11	384.65	4.29	1750.806
1988	83.22	665.79	6.58	2490.062
1989	108.03	864.21	7.56	2634.625
1990	140.20	1402.05	10.85	3486.121

HD: D = 8.75 in, H = 1 in

Year	Areal Capacity (MB)	Volumetric Capacity (MB)	Data Rate (MB/sec)	Total Hard Disk Cost (\$)
1980	28.55	57.11	0.88	1070.463
1981	37.09	74.19	1.02	1142.676
1982	48.19	96.39	1.17	1205.535
1983	62.60	125.19	1.35	1286.591
1984	81.30	162.59	1.56	1358.587
1985	105.58	211.15	1.80	1449.727
1986	137.12	274.25	2.07	1532.472
1987	178.03	356.05	2.39	1620.647
1988	231.12	462.24	2.79	1725.921
1989	300.03	600.06	3.27	1849.890
1990	389.42	778.84	3.83	2003.714

HD: D = 8.75 in, H = 2 in

Year	Areal Capacity (MB)	Volumetric Capacity (MB)	Data Rate (MB/sec)	Total Hard Disk Cost (\$)
1980	28.55	114.21	1.76	2140.926
1981	37.09	148.38	2.03	2285.351
1982	48.19	192.77	2.35	2411.070
1983	62.60	250.38	2.71	2573.182
1984	81.30	325.18	3.13	2717.174
1985	105.58	422.31	3.60	2859.960
1986	137.12	548.48	4.13	2992.415
1987	178.03	712.12	4.71	3115.442
1988	231.12	924.48	5.34	3229.040
1989	300.03	1200.06	6.03	3334.299
1990	389.42	1557.68	6.87	3431.221

A.4 Color CRT Display Supply Model Output

The following section is organized as follows:

- Subsection A.4.1 lists the maximum number of holes per inch of a color CRT metal shadow mask.
- Subsections A.4.2 through A.4.5 present the parameters, costs, and cost per megapixel of 16-inch, 19-inch, 20-inch, and 25-inch color CRT displays. The parameters include the number of pixels per inch, the resolution, the horizontal scanning frequency, and the bandwidth.

A.4.1 Maximum Number of Holes in the Metal Shadow Mask: 1985-1990

Color CRT	
Year	Max-#Holes/inch
1985	55.81
1986	61.19
1987	67.09
1988	73.57
1989	80.66
1990	88.45

A.4.2 16-inch Color CRT Parameters and Cost: 1985-1990

16-inch Color CRT					
Year	#Pixels/inch	Resolution (HxV)	H-Scan (KHz)	Bandwidth (MHz)	Total CRT Cost (\$)
1985	55.81	714x535	50.00	26.79	1033.26
1986	61.58	788x591	55.17	32.61	999.95
1987	67.80	867x650	60.75	39.54	968.75
1988	73.01	934x700	65.42	45.85	948.44
1989	80.25	1027x770	71.90	55.39	919.76
1990	88.07	1127x845	78.91	66.71	892.64

16-inch Color CRT	
Year	Cost/MP (\$)
1985	2700.00
1986	2146.21
1987	1715.07
1988	1447.88
1989	1162.29
1990	936.60

A.4.3 19-inch Color CRT Parameters and Cost: 1985-1990

19-inch Color CRT					
Year	#Pixels/inch	Resolution (HxV)	H-Scan (KHz)	Bandwidth (MHz)	Total CRT Cost (\$)
1985	55.81	848x636	59.38	37.78	1457.06
1986	61.58	935x701	65.52	45.99	1410.29
1987	67.80	1030x772	72.14	55.76	1366.48
1988	73.01	1109x832	77.69	64.66	1338.04
1989	80.25	1219x914	85.38	78.11	1297.77
1990	88.07	1338x1003	93.70	94.08	1259.70

19-inch Color CRT	
Year	Cost/MP (\$)
1985	2700.00
1986	2146.52
1987	1715.57
1988	1448.52
1989	1162.98
1990	937.30

A.4.4 20-inch Color CRT Parameters and Cost: 1985-1990

20-inch Color CRT					
Year	#Pixels/inch	Resolution (HxV)	H-Scan (KHz)	Bandwidth (MHz)	Total CRT Cost (\$)
1985	55.81	892x669	62.50	41.86	1614.47
1986	61.58	985x738	68.97	50.96	1562.71
1987	67.80	1084x813	75.93	61.78	1514.24
1988	73.01	1168x876	81.77	71.65	1482.79
1989	80.25	1283x962	89.88	86.55	1438.22
1990	88.07	1409x1056	98.64	104.24	1396.10

20-inch Color CRT	
Year	Cost/MP (\$)
1985	2700.00
1986	2146.61
1987	1715.72
1988	1448.71
1989	1163.19
1990	937.51

A.4.5 25-inch Color CRT Parameters and Cost: 1985-1990

25-inch Color CRT					
Year	#Pixels/inch	Resolution (HxV)	H-Scan (KHz)	Bandwidth (MHz)	Total CRT Cost (\$)
1985	55.81	1116x837	78.13	65.40	2522.61
1986	61.58	1231x923	86.21	79.62	2442.20
1987	67.80	1355x1016	94.92	96.53	2366.89
1988	73.01	1460x1095	102.22	111.95	2318.18
1989	80.25	1604x1203	112.35	135.24	2248.95
1990	88.07	1761x1321	123.30	162.88	2183.53

25-inch Color CRT	
Year	Cost/MP (\$)
1985	2700.00
1986	2147.01
1987	1716.37
1988	1449.53
1989	1164.08
1990	938.42

A.5 UNIX Supply Model Output

The following section lists, in Subsections A.5.1 and A.5.2, the porting and development-from-scratch time periods and costs of the UNIX operating system, respectively.

A.5.1 UNIX Porting Time Period and Cost: 1980-1990

UNIX: Porting		
Year	Porting Time (Yrs)	Porting Cost (\$)
1980	1.50	210000.00
1981	1.46	198948.32
1982	1.40	184962.68
1983	1.37	175246.32
1984	1.31	163267.11
1985	1.26	151599.47
1986	1.21	141476.31
1987	1.16	132209.98
1988	1.12	123501.21
1989	1.08	115549.42
1990	1.03	107015.41

A.5.2 UNIX Development-from-Scratch Time Period and Cost: 1980-1990

UNIX: Development-from-Scratch		
Year	Development Time (Yrs)	Development Cost (\$)
1980	15.00	2100000.00
1981	14.64	1989483.16
1982	14.02	1849626.76
1983	13.68	1752463.16
1984	13.13	1632671.14
1985	12.55	1515994.73
1986	12.07	1414763.08
1987	11.61	1322099.81
1988	11.17	1235012.09
1989	10.77	1155494.25
1990	10.27	1070154.07

BIBLIOGRAPHY

- [1] Apiki, S. and Diehl, S. (1989) "Upscale Monitors," *Byte*, March, pp. 162-174.
- [2] Arcuri, F. (1989), "Market Survey: CRT Computer Monitors," *Information Display*, June, pp. 10-13.
- [3] Bajorek, C. H. (1989), "Trends in Recording and Control and Evolution of Subsystem Architectures for Data Storage," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 1-(1-6).
- [4] Bell, G. (1988), "Toward a History of Personal Workstations," in **A History of Personal Workstations**, Goldberg, A., Reading, MA: Addison-Wesley Publishing Company.
- [5] Berghof, W. (1989), "Head and Media Requirements for High-Density Recording," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 1-(97-99).
- [6] Bertsekas, D. P. and Tsitsiklis, J. N., **Parallel and Distributed Computation: Numerical Methods**. Englewood Cliffs, NJ: Prentice Hall, Inc., 1989.
- [7] Boschma, B. D., et al. (1989), "A 30 MIPS VLSI CPU," *IEEE International Solid-State Circuits Conference: Digest of Technical Papers*, pp. 82-83.

- [8] Bowater, R. J. (1989), "The IBM Image Adapter/ATM," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 1-(104-108).
- [9] Breen, P. T. (1988), "Workstations-New Environment, New Market," *Information Display*, November, pp. 8-10.
- [10] Brodsky, M. H. (1990), "Progress in Gallium Arsenide Semiconductors," *Scientific American*, February, pp. 68-75.
- [11] Brooke, A., Kendrick, D. and Meeraus, A., **GAMS: A User's Guide**. Redwood City, CA: The Scientific Press, 1988.
- [12] Cates, R. (1990), "Gallium Arsenide Finds a New Niche," *IEEE Spectrum*, April, pp. 25-28.
- [13] Chang, I. F. (1980), "Recent Advances in Display Technologies," *Proceedings of the Society of Information Display*, Vol. 21, pp. 45-54.
- [14] Chesters, M. J. (1989), "A 1 micron CMOS 128 MHz Video Serializer, Palette and Digital-to-Analogue (DAC) Chip," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 1-117.
- [15] Chow, P. (1991), "RISC (Reduced Instruction Set Computing)," *IEEE Potentials*, October, pp. 28-31.
- [16] "Color CRT Prices," provided by David Eccles, *Sony Corporation - Engineering Division*, San Diego, CA, 1991.
- [17] "Computer Confusion: A Jumble of Competing, Conflicting Standards is Chilling the Market," *Business Week*, June 10, 1991, pp. 72-78.

- [18] Conversations with Al Tasch, Professor at the Department of Electrical and Computer Engineering, the University of Texas at Austin, Austin, Texas, 1991.
- [19] Cunningham, J. A. (1990), "The Use and Evaluation of Yield Models in Integrated Circuit Manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, Vol. **3**, **5**, pp. 60-71.
- [20] Dill, F. H. (1982), "Future Trends and Mutual Impact of VLSI and Display," *1982 International Display Research Conference*, pp. 9-10.
- [21] Flores, G. E. and Kirkpatrick, B. (1991), "Optical Lithography Stalls X-Rays," *IEEE Spectrum*, October, pp. 24-27.
- [22] Forsyth, M., Mangelsdorf, S., DeLano, E, Gleason, C., Steiss, D., and Yetter, J. (1991), "CMOS PA-RISC Processor for a new Family of Workstations," *COMPCON Spring '91: Digest of Papers*, pp. 202-207.
- [23] Funk, H. L. (1989), "Information Displays - An Overview and Trends," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 2-(1-7).
- [24] "Gallium Arsenide Chips: Half Way to Paradise," *The Economist*, June 15, 1991, p. 83.
- [25] Garey, M. R. and Johnson, D. S., **Computers and Intractability: A Guide to the Theory of NP-Completeness**. New York, NY: W.H. Freeman and Company, 1979.

- [26] George, J. (1989), "High-Technology Competition between U.S. and Japanese Companies," *Japanese Business Study Program*.
- [27] Ghausi, M. S., **Electronic Devices and Circuits: Discrete and Integrated**. New York, NY: Holt, Rinehart and Winston, 1985.
- [28] Goede, W. F. (1982), "Technologies for High-Resolution Color Display," *1982 International Display Research Conference*, pp. 60-62.
- [29] Goldberg, A., **A History of Personal Workstations**. Reading, MA: Addison-Wesley Publishing Company, 1988.
- [30] Haemer, J. S, McCarron, S. P. and Salus, P. H. (1991), "Trends in UNIX Software," *COMPCON Spring '91: Digest of Papers*, pp. 365-368.
- [31] Hamacher, V. C. , Vranesic, Z. G. and Zaky, S. G., **Computer Organization**. New York, NY: McGraw-Hill Book Company, 1984.
- [32] Hayes, F. (1991), "In Search of Standard Unix," *UnixWorld*, December, pp. 91-92.
- [33] Hennessy, J. L. and Patterson, D. A., **Computer Architecture: A Quantitative Approach**. San Mateo, CA: Morgan Kaufman Publishers, Inc., 1989.
- [34] Hennessy, J. L. and Jouppi, N. P. (1991), "Computer Technology and Architecture: An Evolving Interaction," *Computer*, September, pp. 18-29.
- [35] Hillier, F. S. and Lieberman, G. J., **Introduction to Operations Research**. Oakland, CA: Holden-Day, Inc., 1986.

- [36] Hönig, H. E. (1989), "Superconductivity," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 5-(79-81).
- [37] "HP Workstations Performance Data," provided by Bryan Brademan, *Hewlett Packard - Sales and Marketing Division*, 1991.
- [38] Iki, T. and Werner, K. (1989), "CRTs," *Information Display*, December, pp. 6-7.
- [39] Infante, C. (1988), "Advances in CRT Displays," *1988 International Display Research Conference*, pp. 9-11.
- [40] Infante, C. (1986), "CRT Technology: Progress and Issues," *Proceedings of the Society of Information Display*, Vol. 27, No. 4, pp. 245-248.
- [41] "Intel's 80x86 Data," provided by Todd J. Derr, University of Pittsburgh, Pittsburgh, Pennsylvania, 1991.
- [42] "Intel's Plan for Staying on Top," *Fortune*, March 27, 1989.
- [43] Jain, R. (1991), "Performance Analysis Ratholes or How to Stall a Performance Presentation," *Computer*, June, p. 112.
- [44] Kallfass, T. (1989), "Thin-Film Transistors for Addressing LC-Flat Panel Displays," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 2-(20-23).
- [45] Kendrick, D., Meeraus, A. and Alatorre, J., **The Planning of Investment Programs in the Steel Industry**. Published for the World Bank by the Johns Hopkins University Press, Baltimore and London, 1984.

- [46] Kendrick, D. A. and Stoutjesdijk, A. J., **The Planning of Industrial Investment Programs: A Methodology**. Published for the World Bank by the Johns Hopkins University Press, Baltimore and London, 1978.
- [47] Kernighan, B. W. and Lin, S. (1970), "An Efficient Heuristic Procedure for Partitioning Graphs," *The Bell System Technical Journal*, February, pp. 291-307.
- [48] Kötzle, G. (1989), "VLSI Technology Trends," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 5-(58-62).
- [49] Kryder, M. H. (1989), "Data Storage in 2000-Trends in Data Storage Technologies," *IEEE Transactions on Magnetics*, November, pp. 4358-4363.
- [50] Liebmann, W. K. (1989), "VLSI-the Driving Force for Computer Peripherals," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. P-(16-17).
- [51] Mahon, M. J., Lee, R. B., Miller, T. C., Huck, J. C., and Bryg, W. R. (1986), "Hewlett-Packard Precision Architecture: The Processor," *Hewlett-Packard Journal*, August, pp. 4-21.
- [52] Markoff, J. "Supercomputing's Speed Quest," *The New York Times*, May 31, 1991.
- [53] Marston, A., et al. (1987), "A 32b CMOS Single-Chip RISC Type Processor," *IEEE International Solid-State Circuits Conference: Digest of Technical Papers*, pp. 28-29.

- [54] Martin, A. (1988), "Ruggedized Color CRT Assemblies," *Information Display*, September, pp. 10-13.
- [55] Masterman, H. C. (1988), "Displays for Workstations," *Information Display*, November, pp. 11-14.
- [56] McHaney, R., **Computer Simulation: A Practical Perspective**. San Diego, CA: Academic Press, Inc., 1991.
- [57] Mee, C. D. and Daniel, E. D., **Magnetic Recording Handbook**. New York, NY: McGraw-Hill Publishing Company, 1990.
- [58] " 'Micros' Vs. Supercomputers," *The New York Times*, May 6, 1991.
- [59] Mills, R. (1989), "Why 3D Graphics?," *Computer-Aided Engineering*, March, pp. 50-68.
- [60] Money, S. A., **Microprocessor Data Book**. New York, NY: McGraw-Hill Book Company, 1982.
- [61] "Motorola's 68K Series: Die Sizes," provided by Roy Druian, *Motorola - 68000 Marketing and Applications Division*, 1991.
- [62] "Motorola DRAM Data," provided by Judy Racino, *Motorola - Marketing Communications Division*, 1991.
- [63] Myers, W. (1991), "Five Plenary Addresses Highlight Compcon Spring 91: GaAs Targets 100-MHz-plus Computers," *Computer*, May, pp. 102-104.
- [64] Myers, W. (1991), "The Drive to the Year 2000," *IEEE Micro*, February, pp. 10-13, 68-74.

- [65] Nakanishi, H., Okuda, S., Yoshida, T. and Sugahara, T. (1986), "A High Resolution Color CRT for CAD/CAM Use," *Proceedings of the Society for Information Display*, Vol. **27**, No. **2**, pp. 153-156.
- [66] Ohsaki, T. (1991), "Electronic Packaging in the 1990's - A Perspective From Asia," *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, June, pp. 254-261.
- [67] Opie, R. (1987), "Producing Color Graphic Display," *Control and Instrumentation*, October, 47-48.
- [68] "Panel Gives Chip Outlook," *The New York Times*, May 16, 1991.
- [69] *PC Laptop Computers Magazine*, January, 1992.
- [70] Peterson, J. L. and Silberschatz, A., **Operating System Concepts**. Reading, MA: Addison-Wesley Publishing Company, 1987.
- [71] Phone conversations with Charles Malear, *Motorola - Advanced Microcontroller Division*, 1991.
- [72] Phone conversations with Roger McKee, *Uniform - Vice President of Marketing and Member Services*, 1991.
- [73] Reichl, H. (1989), "Packaging of VLSI Devices," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 5-(63-67).
- [74] Riezenman, M. J. (1991), "Wanlass's CMOS Circuit," *IEEE Spectrum*, May, p. 44.

- [75] Rosch, W. L. (1991), "Mainstream Monitors: What Really Matters," *PC Magazine*, July, pp. 103-186.
- [76] Rosen, B. and Kriz, S. (1988), "Case Study: Developing a 3000-Line Interactive CRT Display," *Information Display*, January, pp. 12-15.
- [77] Salus, P. H. (1991), "UNIX Software Next...," *COMPCON Spring '91: Digest of Papers*, pp. 362-364.
- [78] Schadt, M. (1989), "Electro-Optical Effects, Liquid Crystals and their Application in Displays," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 2-(15-19).
- [79] Seidman, A. H. and Flores, I., **The Handbook of Computers and Computing**. New York, NY: Van Nostrand Reinhold Company, Inc., 1984.
- [80] Shmulovich, J. (1989), "Advanced Technology: Thin Film CRT Phosphors," *Information Display*, March, pp. 17-19.
- [81] Stewart, G. A. (1988), "Multiscan Color Monitors," *Byte*, February, pp. 101-115.
- [82] Stone, H. S. and Cocke, J. (1991), "Computer Architecture in the 1990s," *Computer*, September, pp. 30-38.
- [83] Strum, W. E. (1988), "Trends in Microcomputer Image Processing," *SPIE Proceedings, Vol. 900, Imaging Applications in the Work World*, pp. 3-6.

- [84] "Sun Workstations Pricing History and Performance Data," provided by David Cohn, *Sun Microsystems, Inc. - District Sales Support Division*, 1989.
- [85] "Sun Leads in Workstations," *The New York Times*, January 22, 1990.
- [86] "Sun Workstations Performance Data," provided by Andrea Pusateri, *Sun Microsystems, Inc. - Sales and Marketing Division*, 1991.
- [87] Suntola, T. (1989), "Thin-Film EL-Displays," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 2-(32-35).
- [88] Takata, H. (1989), "Future Trend of Storage Systems," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 1-(7-11).
- [89] Tanksalvala, D., et al. (1990), "A 90MHz CMOS RISC CPU Designed for Sustained Performance," *IEEE International Solid-State Circuits Conference: Digest of Technical Papers*, pp. 52-53.
- [90] Tannas, L. E. Jr., **Flat Panel Displays and CRTs**. New York, NY: Van Nostrand Reinhold Company, Inc., 1985.
- [91] Tannas, L. E. Jr. (1989), "Flat Panel Displays Displace Large, Heavy, Power-Hungry CRTs," *IEEE Spectrum*, September, pp. 35-36.
- [92] Tasch, A., **Class Notes for EE396K, MOS-IC Process Integration**. Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas, 1990.
- [93] Terry, C. (1987), "Refinements in CRT Design Boost Resolution of Color Video Monitors," *EDN*, October 29, pp. 81-84.

- [94] "The Open Software Foundation: A Look at Computing in the 1990s," *COMPCON Spring '91: Digest of Papers*, pp. 369-374.
- [95] Tummala, R. R. (1991), "Electronic Packaging in the 1990's - A Perspective From America," *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, June, pp. 262-271.
- [96] Virgin, L. (1987), "Understanding and Evaluating a Computer Graphics Display," *Information Display*, December, pp. 17-19.
- [97] Waldecker, D. (1991), "Presentation at the University of Texas at Austin: IBM RISC System/6000".
- [98] Weicker, R. P. (1990), "An Overview of Common Benchmarks," *Computer*, December, pp. 65-75.
- [99] Wessely, H., Fritz, O., Horn, M., Klimke, P., Koshnick, W. and Schmidt, K. H. (1991), "Electronic Packaging in the 1990's - A Perspective From Europe," *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, June, pp. 272-284.
- [100] Wilczynski, J. (1989), "Low Temperature CMOS VLSI Technologies," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 5-(73-78).
- [101] Wood, R. (1990), "Magnetic Megabits," *IEEE Spectrum*, May, 5, pp. 32-38.
- [102] Wurtz, J. E. (1989), "The Not-So-Amazing Survival of the CRT," *Information Display*, September, pp. 5-6,18.

- [103] Yetter, J., Forsyth, M., Jaffe, W., Tanksalvala, D. and Wheeler, J. (1987), "A 15MIPS 32b Microprocessor," *IEEE International Solid-State Circuits Conference: Digest of Technical Papers*, pp. 26-27.
- [104] Yoffie, D. B. and Wint, A. G. (1987), "the Global Semiconductor Industry, 1987," *Harvard Business School, Case #9-388-052*.
- [105] Zeidler, H. C., (1989), "Intelligent Access to Mass Memories," *IEEE Proceedings, COMPEURO'89: VLSI and Computer Peripherals*, pp. 1-(27-31).

VITA

Walid Rachid Touma was born in Kab-Elias, the Bekaa Valley, Lebanon, on April 2, 1965, the son of Rachid Tanios Touma and Laure Layoun Touma. In Lebanon, he attended the Athenee de Beyrouth High School, Rabieh. In 1984, he joined the engineering program of the University of Texas at Austin. He received the degree of Bachelor of Science, with High Honors, in Electrical and Computer Engineering in May 1987. In September 1987, he joined the Graduate School of the University of Texas at Austin and earned the degree of Master of Science in Electrical and Computer Engineering in December 1989. The title of the thesis is "Clustering/Partitioning Algorithms and Comparative Analysis." In January 1990, he assisted his supervisor, Professor Martin L. Baughman, and five other faculty at the University of Texas at Austin in obtaining a grant from DARPA to fund the "Workstation Project" on which he and Professor Baughman began working during the summer of 1989.

Permanent address: P.O. Box 49843
Austin, Texas 78765.

This dissertation was typeset¹ with \LaTeX by the author.

¹The \LaTeX document preparation system was developed by Leslie Lamport as a special version of Donald Knuth's \TeX program for computer typesetting. \TeX is a trademark of the American Mathematical Society. The \LaTeX macro package for The University of Texas at Austin dissertation format was written by Khe-Sing The.