1 Introduction

> Ideally one would desire an indefinitely large memory capacity such that any particular ... word would be immediately available. ... We are ... forced to recognize the possibility of constructing a hierarchy of memories, each of which has greater capacity than the preceding but which is less quickly accessible.
>
> A. W. Burks, H. H. Goldstine, and J. von Neumann
> 1946 [73, p. 402]

For several decades, high-performance computer systems have incorporated a memory hierarchy [73]. Because central processing unit (CPU) performance has risen much more rapidly than memory performance since the late 1970s, modern computer systems have an increasingly severe performance gap between CPU and memory. Figure 1 illustrates this trend, and Figure 2 shows that performance increases are happening at both the high and low ends of the market. Computer architects have attempted to compensate for this performance gap by designing increasingly complex memory hierarchies.

> Clock increases in speed do not exceed a factor of two every five years (about 14%).
>
> C. Gorden Bell
> 1992 [12, p. 35]

> ... a quadrupling of performance each three years still appears to be possible for the next few years. ... The quadrupling has its basis in Moore's law stating that semiconductor density would quadruple every three years.
>
> C. Gorden Bell
> 1992 [12, p. 40]

> ... (Others) suggest that the days of the traditional computer are numbered. ... Today it is improving in performance faster than at any time in its history, and the improvement in cost and performance since 1950 has been five orders of magnitude. Had the transportation industry kept pace with these advances, we could travel from San Francisco to New York in one minute for one dollar!
>
> John L. Hennessy and David A. Patterson
> 1990 [73, p. 571]

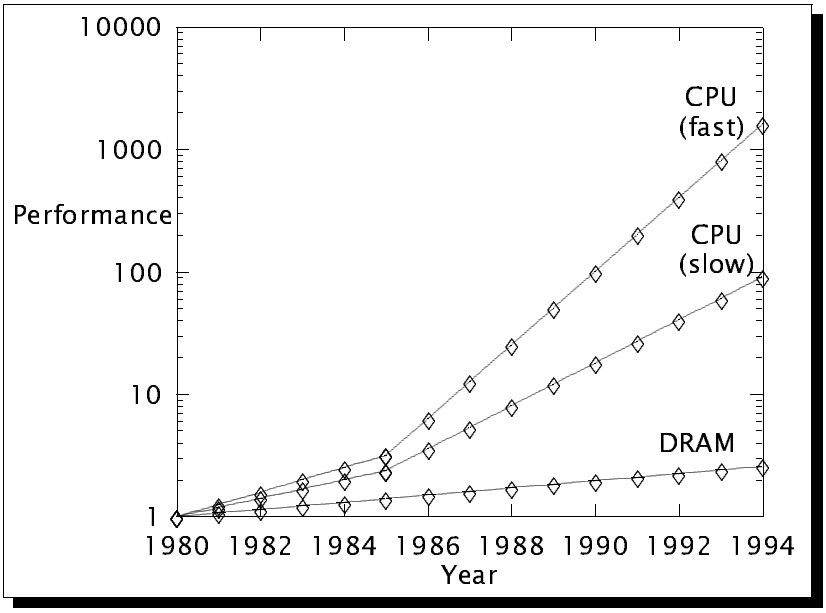

Figure 1: **CPU and memory performance.** This drawing uses 1980 as a baseline. Memory speed (dynamic random-access memory, DRAM) is plotted with an annual 7% increase. The slow CPU line grows at 19% annually until 1985, and at 50% annually since then. The fast CPU line rises at 26% annually until 1985, and at 100% annually since then. The data is taken from [73, Fig. 8.18, p. 427], but extended beyond 1992.

The existence of a memory hierarchy means that a few well-behaved programs will perform almost optimally on a particular system, but alas, most will not. Because the performance difference between the extremes of good and bad behavior can be several orders of magnitude, it is important for programmers to understand the impact of memory access patterns on performance.

Fortunately, once the issues are thoroughly understood, it is usually possible to control memory access in high-level languages, so it is seldom necessary to resort to assembly-language programming, or to delve into details of electronic circuits.

The purpose of these notes is to give the reader a description of the computer memory hierarchy, and then to demonstrate how a programmer working in a high-level language can exploit the hierarchy to achieve near-optimal performance.

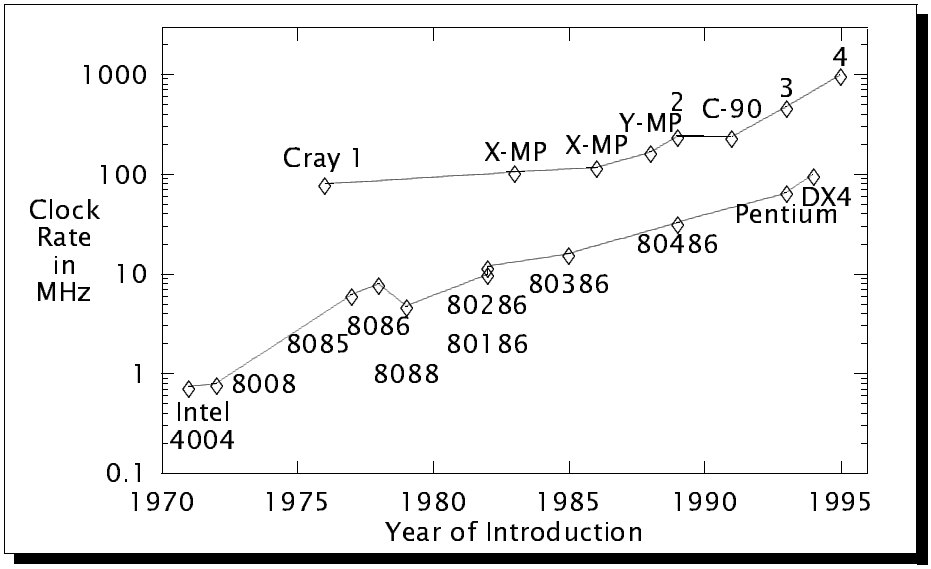

Figure 2: **Rising clock rates of supercomputers and microcomputers.** In microcomputers, each CPU model is generally available in a range of clock rates; the figures shown here are the fastest at the time of introduction.
Chips are fabricated on silicon disks, or wafers, about 15 cm in diameter, with up to several hundred chips per wafer. When they are subsequently cut from the wafer and packaged in ceramic or plastic and tested, some are found to work at higher clock rates than others; the faster ones command a higher price. For photographs and cost details, see [73, pp. 53–66].
Supercomputer clock rates are fixed because of the highly-optimized, and delicate, design balance between CPU, memory, buses, and peripherals.

2 Acronyms and units

Like any other technical field, computing has its own jargon, and numerous acronyms. We define technical terms as we encounter them, italicizing their first occurrence, and include them in the index at the end to make it easy to find the definitions. All of the authors of cited works and displayed quotations are indexed as well, so you can find your way from the bibliography back into the text where the citation is discussed.

Manufacturers commonly designate models of computers and their peripherals by numbers. Some of these, like 360 and 486, become so well known that they are commonly used as nouns, without naming the company. Computer products are indexed both by vendor, and by model number.