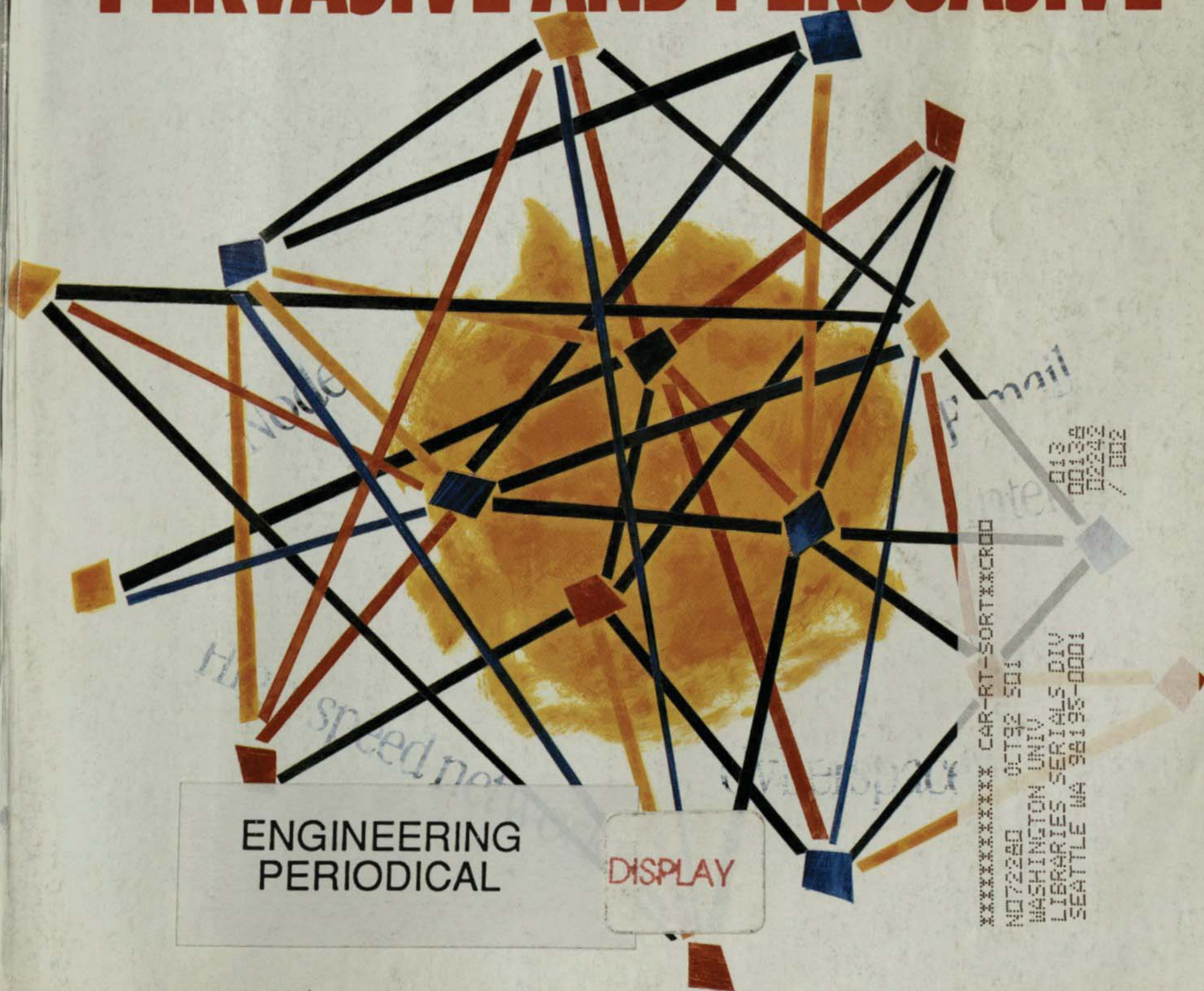
HIGH-SPEED DRAMS, P. 34

IEEE SPECTRUM

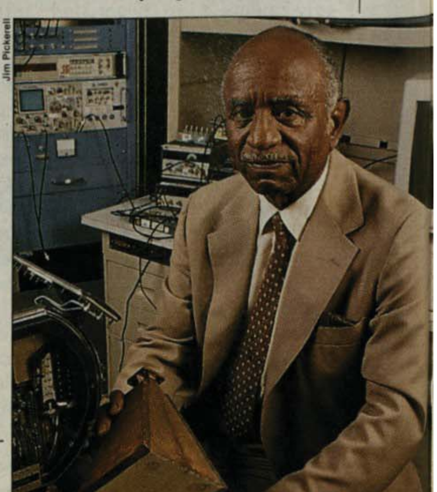
E-MAIL
PERVASIVE AND PERSUASIVE



THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

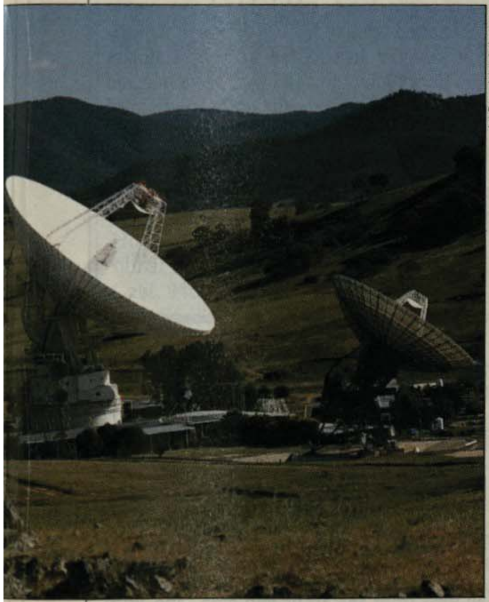
OCTOBER 1992 Volume 29 Number 10

IEEE SPECTRUM



Jim Pickerell

SYSTEMS



Jet Propulsion Laboratory

How a system is created, designed, and built blends art and engineering. NASA's Deep Space Network is the product of such a process. Here, great antennas peer into space from its Canberra, Australia, station. The article describes the architecting process and provides additional examples.

SPECTRAL LINES

"Scientific management" gives ambiguous guidance today. Nevertheless, managers have to select among traditional, perhaps outmoded, concepts and contemporary techniques, perhaps fads, to develop a self-consistent management process.

Cover: Electronic messages travel over a global network unbounded by time zones, distance, or political entities in Gus Sauter's conceptual illustration. With e-mail, an engineer can communicate with a colleague halfway around the world as easily as with a co-worker down the hall. The technology, still in its infancy, is changing society as well as business. *Spectrum*'s Special Report on e-mail begins on p. 22.

IEEE SPECTRUM (ISSN 0018-9235) is published monthly by The Institute of Electrical and Electronics Engineers, Inc.

345 East 47th St., New York, N.Y. 10017, U.S.A. Cable address: ITRIPLEE. Telex 236-411. Fax: 212-705-7453. E-mail: ieeespectrum.

ANNUAL SUBSCRIPTIONS. IEEE members: $11.00 included in dues. Nonmembers: $29.95. Libraries/institutions: $139.

SINGLE COPIES. Members: $8. Nonmembers: $16.

MICROFICHE SUBSCRIPTIONS. Members: $16. Nonmembers and libraries: $139.

POSTMASTER: Please send address changes to *IEEE Spectrum*, c/o Coding Department, IEEE Service Center, 445 Hoes Lane, Box 1331, Piscataway, N.J. 08855. Second Class postage paid at New York, N.Y., and additional mailing offices. Canadian GST #125634188.

Printed at 8649 Hacks Cross Rd., Olive Branch, Miss. 38654.

IEEE Spectrum is a member of the Audit Bureau of Circulations, the Magazine Publishers of America, and the Society of National Association Publications.

Staff

Advisory Board

Editorial Board

Fast computer memories

Designers are searching for new DRAM technologies to reduce memory access time and so unleash computer performance

If the price-to-performance ratio of computer systems is to keep improving, the gap in speed between processors and memory must be closed. Processors perform at their peak only when the flow of instructions and data from memory is fast and unfaltering.

An ever-flowing stream is particularly necessary to reduced-instruction-set computing (RISC) processors, which have become very popular during the last few years. A heavily pipelined RISC processor can execute an instruction every clock cycle, demanding a lot of the memory system. Both superscalar processors, with their multiple functional units, and multiprocessor machines make even greater demands on memory systems.

Nor is the central processing unit (CPU) the only consumer of memory bandwidth. Computers now are expected to be easier to use and more capable than their predecessors, and some of the new capabilities will require speedier memory. Examples include the rapid display of high-resolution graphics in true color, the recognition of speech and handwriting, recording and playing back video and audio, and, ultimately, support of a virtual-reality environment. The machines may also need buffer memory for messages moving over the multigigabit-per-second networks expected in the near future.

These lofty I/O ambitions all involve processing and moving large amounts of data. On top of even faster CPUs, they will strain both memory capacity and memory bandwidth. But the accepted dynamic RAM (DRAM) architectures and solutions have been pushed to their limits. A basic change in architecture seems the only way to obtain an urgently needed increase in memory speed.

The need for change has struck a number of chip makers, because innovative architectures distinguish a variety of recent high-speed DRAMs, which go by such names as synchronous, cached, and Rambus DRAMs. The newcomers may be usefully surveyed from a system perspective, to see how they may solve design problems, particularly with regard to main memory.

Ray Ng Sun Microsystems Inc.

MEMORY LINE. Till now, in the familiar stored-program computer described by von Neumann, the processor has been connected directly to memory (as well as to input/output). From this model, a hierarchical memory system has evolved in which a little, very fast memory is placed very close to the processor and fed by lots of slower memory farther away from the processor [Fig.1]. This hierarchy, which is used in almost all computer systems today, reflects one of computer design's truisms, "fast memory is expensive and slow memory is cheap."

At the first level of the hierarchy are the processor's internal registers. Access to these registers is very fast because they are on the processor chip. However, their number is limited by the available chip area, or "real estate."

At the second level, between the processor and slower main memory, is a cache—a small, very fast memory. The cache is loaded with copies of those blocks of data stored in main memory that the processor is most likely to want for the operation it is currently performing. (Typically, the block size ranges from 16 to 64 bytes.)

> Processors can perform their best only if the supply of instructions and data is fast and unfaltering

If the processor finds the data it wants in the cache (referred to as a cache hit), then to the processor it will look as if main memory is as fast as cache. But if the processor does not find what it needs in cache (a cache miss), then the block containing the missing data must be brought in from main memory, slowing down the system.

Caches usually provide a speedup because they exploit a general characteristic of programs: locality in space and time. Spatial locality indicates that if a location in memory is accessed, then others nearby will probably be accessed soon; temporal locality means that if a location in memory is accessed once, then it will probably be accessed again soon.

One problem with caches is that, in order to be effective, they require very fast RAMs that run at about the same speed as the processor; and while static RAMs (SRAMs) can deliver the required speed, they are expensive. Also, caches must keep track of which memory blocks are in the cache and what their state is, and therefore require a special controller and a tag memory that add complexity and take up precious board real estate. All the same, caches are popular.

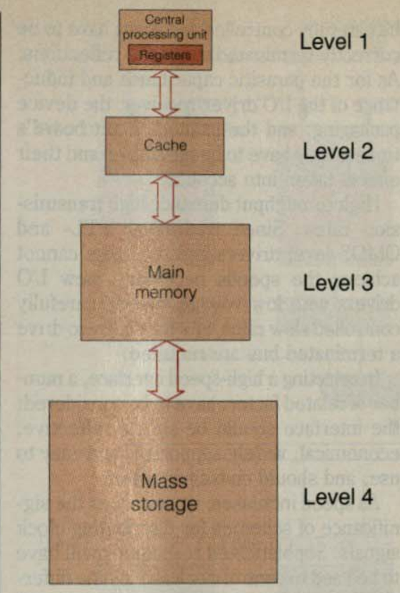
It is possible, too, to build systems with more than one level of caching, using on- and off-chip memory. Many modern processors have on-chip caches, for both program instructions and data, that are closer than an external cache and so faster to access. But like the number of registers, the caches have to be small because chip real estate is limited and in many systems they are supplemented with an external cache. The internal cache is referred to as first-level cache and the external as second-level.

The third level of the hierarchy is main memory itself. Main memory is used to store programs and data, and as a source of input and destination for output. Typically, this memory is much larger than cache and is constructed of DRAMs, which are slower than the SRAMs but also less expensive.

The fourth level of the hierarchy is mass storage. Today magnetic-disk storage is ubiquitous. It is used to implement a technique called virtual memory, which fools the processor into thinking the main memory is much larger than is the case. With virtual memory, the processor's address space is divided into blocks of fixed size, called pages. Pages are much larger than cache blocks, usually 4–8K bytes.

Disk bears much the same relationship to main memory as main memory does to cache. Pages are called from disk and placed in main memory when they are needed or returned to disk when they are not. As with cache, the principle of locality is basic. To maintain order in the system, a memory management unit (MMU) keeps track of which pages are in main memory and what

0018-9235/92/$3.00©1992 IEEE



[1] In most computer systems today, the total memory consists of a hierarchy of media. Passing from the top to bottom of the hierarchy, the density of the medium (the amount of data it can store per unit area) increases, while its speed in delivering data and its cost per bit decrease. Some new dynamic RAM (DRAM) technologies aim at simplifying this hierarchy by speeding up main memory to the point where the need for a separate, external cache is moot.

their status is.

As with a cache miss, performance falls off whenever a page is not in main memory when needed (a page fault). The penalty is, however, higher because mechanical disks are much slower than semiconductor main memory. But disks are very cheap, in terms of cost per bit, and can store vast amounts of data; hard disks today commonly store hundreds of megabytes, and the use of magneto-optical and optical discs capable of storing gigabytes is growing.

Strictly speaking, there is a fifth level of storage, for data that will not be used for an extended period of time or whose importance demands its preservation. This archival storage often consists of magnetic tape; of course, removable magnetic and optical discs are also used for long-term storage of programs and data. This storage level has no impact on run-time system operation and so will be ignored for now.

MAIN EVENT. Main memory is almost always implemented using DRAMs, which in both speed and price lag behind the SRAMs generally used for cache. DRAMs use one transistor-capacitor pair, referred to as a cell, to store one bit of information, while SRAMs use a four- or six-transistor flip-flop to store each bit. Because each DRAM cell is very small, DRAMs can be made very dense; the densest DRAM now available is a 16M-bit part, while the densest SRAM is about 4M bits. The per-bit cost of SRAM, depending on its speed, is 5–10 times that of DRAM.

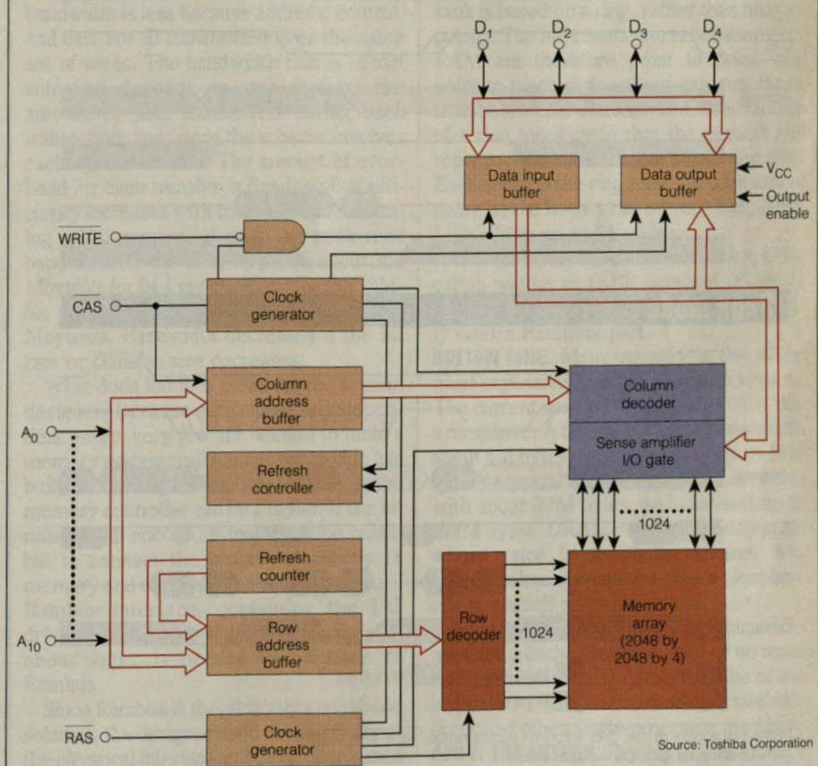
However, because the charge leaks away from the DRAM cell's capacitor, it must be restored by periodic refreshing. Also, the act of reading a DRAM involves transferring and sensing mere dribbles of charge; since each read operation disturbs the cell contents, it, too, requires that the data read be restored. For these reasons, DRAMs are not especially fast.

A DRAM is built as a square or rectangular array of cells [Fig. 2]; to read or write data, the processor sends an address to the DRAM, which typically it multiplexes, supplying first the row address, then the column address. For currently available DRAMs, the time it takes a row address to access a cell is about 40–80 ns, for a column address, about 20–40 ns, and the precharge time is about 30–50 ns. Thus the cycle time (the minimum amount of time between memory accesses by the processor) is about 110–150 ns. In contrast, the cells in small CMOS SRAMs may be accessed every 8 ns; larger SRAMs have a longer access time of about 15 ns.

To raise their operating speed, DRAMs have a special operating mode that takes advantage of their internal row-and-column structure, known as page mode. In this mode, when an entire row (or a chip page, but not to be confused with the virtual memory page) is read into the sense amplifiers, the user can keep the row active and merely change column addresses to access all the data. As long as the accesses remain in the page, the DRAM can work faster. For current DRAMs, the page-mode cycle time (or minimum time between column addresses) is about 40–50 ns [Fig. 3].

SPEEDING UP MAIN MEMORY. A main memory system has three crucial attributes: size, latency, and throughput. Size is affected by density, or the number of bits that can be packed into a given area; the higher the density, the better. Latency is how long it takes for data to be delivered after it has been requested, and is closely related to a DRAM's access time; the shorter the latency, the faster the DRAM. Throughput is a measure of how much data can be delivered in a given period of time, and is closely related to the DRAM cycle time; a higher throughput means that a DRAM delivers more data per time interval. While the density of DRAMs has been quadrupling roughly every three years, neither their access nor their cycle times have improved as rapidly. Improving the latency and throughput of main memory is the focus of attention among memory system designers.

Page mode may reduce latency and increase throughput, but only if there is a lot of sequentiality in the memory reference stream; for multiprocessor systems, this is not true. Caches do a good job of isolating the processor from relatively sluggish main memory, but there is only so much a cache



[2] In a typical DRAM, such as the 4M-by-4-bit Toshiba Corp. part, the actual memory array in the lower right is accessed through paired successive row and column addresses. The row address causes the data in the row to be read into the sense-amplifier I/O gate from which the column address decoder selects the 4-bit word, or nibble, to be placed in the data-out buffer.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | **sales@docketalarm.com** | **1-866-77-FASTCASE**

fastcase®
Smarter legal research.