АКАДЕМИЯ НАУК СССР
СИБИРСКОЕ ОТДЕЛЕНИЕ
А В Т О М Е Т Р И Я
№ 1 1990

ОПТИЧЕСКИЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 681.327.5

В. Е. БУТТ, Н. Н. ВЬЮХИНА, В. И. КОЗИК,
Т. Н. МАНТУШ, Б. Н. ПАНКОВ, Е. Ф. ПЕН,
В. Н. ПИОТТУХ-ПЕЛЕЦКИЙ, П. Е. ТВЕРДОХЛЕБ
(Новосибирск)

ПОИСК СОЕДИНЕНИЙ ПО ФРАГМЕНТАМ СТРУКТУРНЫХ ФОРМУЛ В ГОЛОГРАФИЧЕСКОЙ ПАМЯТИ

Введение. Поиск соединений, структурные формулы которых содержат заданный фрагмент,— одна из наиболее трудоемких задач обработки химической информации [1].

Для размещения баз данных (БД) используются в основном накопители на магнитных или оптических дисках с последовательным (побитовым) способом записи-чтения данных. Запросы к содержимому БД требуют передачи значительных объемов информации из внешней в оперативную память ЭВМ для обработки, что существенно ограничивает скорость поиска. Кроме того, для кодирования-декодирования изображений структур необходимы дополнительные вычислительные затраты.

Голографическая память (ГП) со страничной организацией данных обеспечивает хранение и считывание информации двумерными массивами размером $N \times N$. Тот факт, что считываемый из ГП массив данных представлен в оптическом виде, создает естественные предпосылки для его обработки параллельными оптико-электронными (ОЭ) средствами. При этом информация может обрабатываться непосредственно в памяти без ввода просматриваемого массива в ЭВМ. Ввиду высокой плотности упаковки данных такая память имеет довольно большую емкость при относительно небольших размерах. ГП документального типа [2], обеспечивающая хранение изображений в их естественном (т. е. не кодированном) виде, позволяет производить быструю выдачу документов большой размерности.

Таким образом, параллельная ГП в сравнении с магнитной или оптической побитовой позволяет:

обеспечить высокую скорость выдачи данных ($10^8 - 10^9$ бит/с на один порт);

реализовать оптическим способом параллельный ввод массива данных в ОЭ-спецпроцессоры, которые, в частности, могут быть выполнены в виде однокристальных фоточувствительных БИС или СБИС;

использовать свои преимущества в емкости памяти для ускорения поиска (например, путем преобразования сложных, но «экономных» (с точки зрения затрат емкости памяти) алгоритмов вычислений в более простые, но «неэкономные»);

решать задачи поиска изображений с применением аналоговых методов в ГП документального типа.

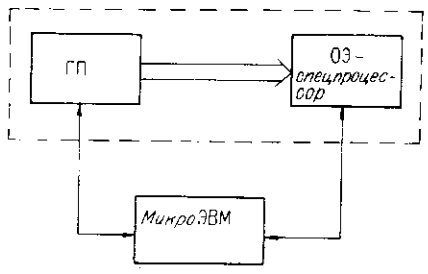

Рис. 1

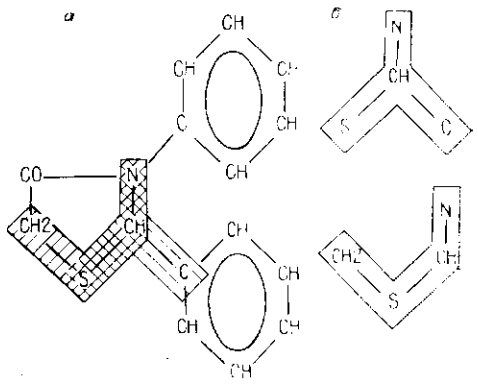

Рис. 2

Блок-схема ОЭ-системы поиска представлена на рис. 1. Поток данных с ГН оптическим способом вводится в ОЭ-спецпроцессор. Данные о химических соединениях представляются в ГН в страничном виде так, что поиск сводится к выполнению спецпроцессором параллельных логических операций. В ЭВМ передаются лишь результаты поиска. Такая система позволяет обеспечить высокую скорость обработки информации, определяемую темпом вывода данных с ГП при использовании внешней ЭВМ низкого быстродействия.

Цель статьи — изложить опыт организации базы данных химических соединений в системе параллельная голографическая память — фотоэлектронный БИС-процессор, обеспечивающей быстрый поиск соединений по фрагментам структурных формул. Рассмотрены использованные в ГП способы представления информации о химических соединениях в виде двумерных массивов данных и параллельные ОЭ-методы ассоциативного поиска данных.

Схема базы данных. Структурная формула химического соединения представляется в виде двумерного графа, в вершинах которого изображены химические элементы, а в ребрах — химические связи. Пример такого графа для соединения 2,3-дифенил-4-тиазолидинол показан на рис. 2, *а*. Фрагментом структурной формулы является некоторый набор связанных вершин данного графа. Так, на рис. 2, *а* могут быть выделены восемь типов 2-вершинных фрагментов (CO—N, CO—CH2, N—C, CH—CH и т. д.), 12 типов 3-вершинных (CO—N—CH, CO—N—C, N—CH—C, N—C—CH и т. д.), 19 типов 4-вершинных (рис. 2, *б*) и т. д. В общей сложности структурные формулы органических соединений содержат в среднем ~100 фрагментов с числом вершин до семи.

Традиционно поиск производится путем сравнения фрагмента, предъявленного в запросе, со всеми однотипными фрагментами набора соединений, выделяемых в БД при помощи первичных дескрипторов, характеризующих определенные структурные особенности в соединениях. Однако ввиду больших объемов БД химических соединений (например, автоматизированный банк данных STN INTERNATIONAL [1] содержит описание более 7×10^6 соединений) при помощи существующих первичных дескрипторов возможно выделять только довольно крупные наборы соединений, а дальнейшая их обработка требует больших вычислительных затрат. Это побудило нас использовать при поиске описания структурных формул в виде их полных фрагментных составов, начиная от 2-вершинных и кончая 7-вершинными. Описания структурных формул представляются в инвертированном виде. В этом случае каждому фрагменту ставится в соответствие список соединений, в состав структурных формул которых он входит.

Очевидно, что такое описание является избыточным, так как фрагменты с меньшим числом вершин могут многократно фигурировать в описаниях более крупных фрагментов, а каждое соединение будет вхо-

дить в списки нескольких фрагментов. Это требует дополнительного объема памяти, однако, как будет показано далее, позволяет довольно эффективно сузить диапазон поиска и значительно упростить алгоритмы обработки данных и тем самым уменьшить время поиска соединений.

Будем исходить из того, что БД содержит информационную и поисковую области. Информационная область включает имеющийся набор описаний соединений, а поисковая — описания структурных формул на основе их фрагментных составов.

В описания соединений включены: регистрационные номера соединений, названия, брутто-формулы, структурные формулы, молекулярные веса; в описания фрагментов — их регистрационные номера, структурные формулы, а также данные о регистрационных номерах соединений, в состав формул которых входит каждый из представленных фрагментов. В основу схемы БД положена реляционная модель. БД представляется совокупностью плоских файлов (таблиц), содержащих данные о различных свойствах соединений. Ключом файла ФРАГМЕНТ, содержащего описания фрагментов, является описание структурной формулы фрагмента, а файла СТРУКТУРА, содержащего описания соединений,— регистрационный номер соединения. В запросе задается структурная формула фрагмента, по которой определяются его номер в БД (при наличии в БД описания идентичного фрагмента), а также количество и список регистрационных номеров искомых соединений. Далее по найденным регистрационным номерам извлекается любая из имеющейся в БД информации о соединении.

n	L_n	q_n	q_r	R_n
2	189	12	15	89
3	744	72	14	326
4	2 383	300	5,5	410
5	5 318	835	3	499
6	9 844	2156	2	616
7	16 516	4653	1,7	878

Примечание. n — число вершин фрагмента; L_n — количество типов фрагментов на 10^3 соединений; q_n — — количество голограмм, содержащих коды фрагментов; q_r — количество голограмм, содержащих списки регистрационных номеров соединений; R_n — — среднее количество номеров в списке соединений.

Кодирование фрагментов структурных формул и формирование страниц данных. Для ускорения поиска кодирование описаний фрагментов выполняется таким образом, чтобы проверку идентичности описаний фрагментов (в запросе и в БД) можно было осуществить с помощью алгоритмов, не требующих больших вычислительных затрат.

Для кодирования фрагментов первоначально составляются таблицы кодов вершин и связей. Все возможные типы вершин и связей размещаются в заданном порядке и нумеруются (число типов вершин 250, связей — 12). Кодом вершины или связи является двоичное представление ее порядкового номера в таблице. Для кодирования типа вершины отводится 1 байт, для кодирования типа связи — 1/2 байта.

При составлении вектора описания фрагмента коды вершин в двоичном представлении (N_i) в порядке возрастания соответствующих им чисел выстраиваются в цепочку, образуя вектор длиной n байт, где n — число вершин в составе данного фрагмента. Далее следуют коды связей между вершинами. На каждую возможную связь (между произвольными l- и k-й вершинами) отводится 1/2 байта: первые $(n-1)/2$ байта на связи первой вершины, далее $(n-2)/2$ на связи второй (исключая первую) и т. д.; всего по числу сочетаний $(n-1)n/2$ позиций или $(n-1)n/4$ байта. Таким образом, каждой паре вершин отводится заданная позиция (1/2 байта) в векторе описания, где указывается код связи (T_{kl}), если таковая существует, или нуль, если данные вершины не связаны между собой. Длина полученного вектора определяется числом вершин фрагмента: $P_n = (n+3)n/4$ байта.

Проверку идентичности фрагментов можно осуществить за один такт сравнением векторов на совпадение, для этого необходимо, чтобы вектор

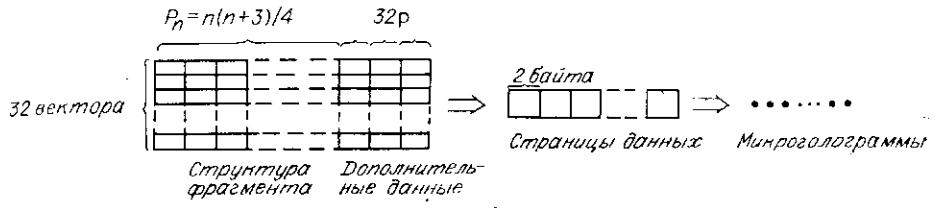

Рис. 3

описания однозначно определялся структурой фрагмента. При наличии однотипных вершин коды их располагаются произвольно, что приводит к изменению вектора кодов связей. Поэтому полученный вектор будет иметь однозначный вид только в том случае, если все вершины, входящие в его состав, различны. В противном случае производится нормализация вектора путем перенумерации однотипных вершин. Выбирается такая их последовательность, при которой двоичное слово, соответствующее вектору описания фрагмента, имеет максимальное значение.

При вводе данных в ГП векторы структур фрагментов дополняются полями номера фрагмента, а также номеров соединений и их количества, в состав структурных формул которых входит этот фрагмент. Затем полученные векторы компонуются в физические двоичные страницы (двумерные массивы). Для обеспечения контроля ошибок в процессе поиска и стабильной мощности сигнального пучка при регистрации голограмм данные представляются в парафазном коде (каждый разряд a_i вектора дополняется инверсным ему, т. е. заменяется на $a_i\bar{a}_i$). Пример компоновки страниц данных поисковой области приведен на рис. 3. Вектор структуры фрагмента дополняется нулями до величины (Q_n), кратной 32. Затем 32 полных вектора (длиной $Q_n + Q_c$, где Q_c — длина вектора, дополняющего описание фрагмента, также кратная 32) формируются в виде матрицы размером $(Q_n + Q_c) \times 32$, которая делится на $(Q_n + Q_c)/32$ субматриц размером 32×32, составляющих группу страниц данных. Таким образом, каждая страница данных содержит 16 парафазных разрядов 32 различных векторов.

Алгоритм поиска. Для ускорения поиска в файлах используются индексирование и упорядочение по первичным ключам. Файл ФРАГМЕНТ упорядочен по значениям поля описания структурных формул фрагментов и имеет, как показано на рис. 4, двухуровневый индекс. Уровню 1

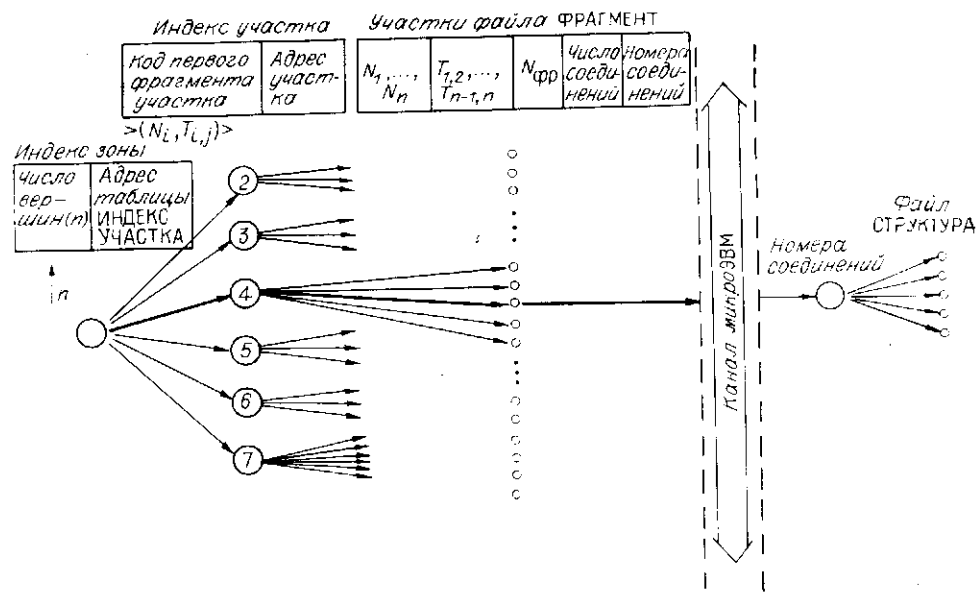

Рис. 4

соответствует ИНДЕКС ЗОНЫ. Файл ФРАГМЕНТ разделен на шесть зон для хранения описаний 2-, 3-, ..., 7-вершинных фрагментов. ИНДЕКС ЗОНЫ служит для определения зоны поиска по числу вершин предъявленного фрагмента. Уровню 2 соответствует ИНДЕКС УЧАСТКА, который представляет собой упорядоченную таблицу значений первичного ключа. Каждое из этих значений содержит наибольшее значение первичного ключа среди всех записей указанного участка. С каждым значением ключа в индексе связан указатель соответствующего участка. Размеры участка и индекса выбираются из соображений минимизации времени поиска.

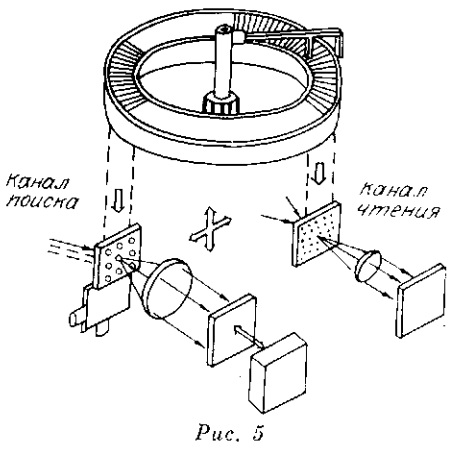

Рис. 5

Схема поиска представлена на рис. 4. Предъявленный для поиска фрагмент кодируется по аналогии со структурами фрагментов БД. По числу вершин фрагмента в ЭВМ, используя ИНДЕКС ЗОНЫ, вычисляются номер зоны, которая может содержать описание данного фрагмента, и адрес соответствующей таблицы ИНДЕКС УЧАСТКА. В результате поиска по упорядоченной таблице ИНДЕКС УЧАСТКА определяется адрес участка памяти (УП), где проводится сравнение кода предъявленного фрагмента и ключей файла ФРАГМЕНТ (значений поля описания структурных формул фрагментов БД). Если в БД обнаружен фрагмент, идентичный предъявленному, в рамках того же участка производится считывание информации о фрагменте.

В процессе поиска, таким образом, выполняются операции считывания данных из ГП, поиска по совпадению в рамках УП и по условию «больше — меньше» в таблицах ИНДЕКС ЗОНЫ и ИНДЕКС УЧАСТКА.

Эксперимент. Цель эксперимента — подтверждение возможности и оценка эффективности предложенного алгоритма поиска данных в системе ГП — ОЭ-спецпроцессор (см. рис. 1).

В качестве ГП использована экспериментальная голографическая память емкостью 1 Гбайт [3]. ГП содержит 300 модулей памяти (МП), на каждом из которых в виде микроголограмм (МГ) размещается 320×320 страниц кодированных данных (32×32 бит). Модуль памяти разбит на 20 полей высотой по 16 МГ. Каждое поле разделено на две части по 8 МГ (основное поле и дублирующее) и включает 320 дорожек. Выбор дорожки осуществляется путем пошагового механического перемещения МП с временем перехода 20 мс, а выбор МГ на дорожке — акустооптическим дефлектором с временем переключения 50 мкс. Вначале проводится чтение МГ с основной части дорожки, а в случае обнаружения ошибки дефлектор переключается на МГ, находящуюся в дублирующей зоне.

Модули в ГП размещаются внутри кассеты, представляющей собой набор радиально ориентированных ячеек (рис. 5). Модуль выбирается механическим захватом и вводится в один из двух каналов — чтения или поиска данных. Во втором канале встроен ОЭ поисковый спецпроцессор, позволяющий проводить ассоциативный поиск параллельно по странице данных.

Спецпроцессор создан на основе однокристального фотоматричного ассоциативного ЗУ (ФМАЗУ) размером 32×36 элементов [4]. ФМАЗУ выполняет следующие микрооперации: 1) параллельная электронная запись 36-разрядного слова по любой группе адресов (строк ФМАЗУ); 2) параллельное оптоэлектронное преобразование страницы данных размером 32×36 бит; 3) хранение записанной информации; 4) электрическое считывание из ФМАЗУ двоичного 36-разрядного слова по заданному

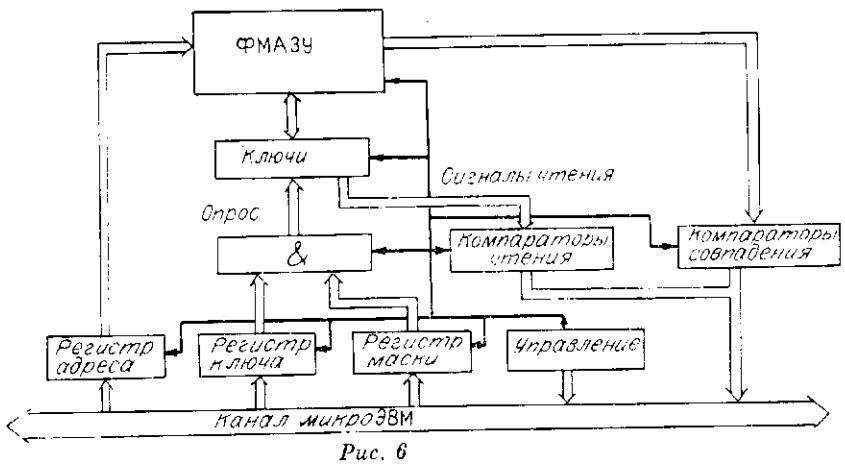

Рис. 6

адресу; 5) параллельное сравнение (с точностью до замаскированных разрядов) входного слова с данными, записанными в ФМАЗУ.

ФМАЗУ представляет собой двумерную решетку фоточувствительных ячеек, имеющих память и образующих матрицу из 32 36-разрядных слов (в данной работе были использованы 32 из 36 разрядов ФМАЗУ), объединенных системой ортогональных разрядных шин записи и выборки в парафазном коде, адресных шин и шин совпадения. Каждая шина адреса, совпадения и разрядная шина имеют отдельный внешний вывод.

Структурная схема управления ФМАЗУ изображена на рис. 6. В режиме поиска проводится параллельное по странице данных сравнение заданных разрядов вектора запроса с соответствующими разрядами векторов (строк) страницы. Страница проецируется в плоскость фотоприемных элементов ФМАЗУ, а соответствующие 2 байта предъявленного в запросе вектора загружаются в регистр ключа. В регистр маски вводится 32-разрядный вектор, определяющий разряды, по которым производится сравнение. Полученное слово опроса (содержимое регистра ключа с выборочно замаскированными разрядами) подается на разрядные шины ФМАЗУ. На шинах совпадения возникают сигналы, пропорциональные числу несовпавших битов слова опроса и каждого из векторов страницы (вычисляется набор функций $F_h = \sum_{i=1}^{32} (a_{hi}\bar{b}_i + \bar{a}_{hi}b_i)$ для 32 векторов, где h — номер строки страницы данных, i — номер разряда, $\{a_i\}_h$ — вектор ключа, b — вектор запроса).

Поиск значений «больше — меньше» реализуется путем поразрядного сравнения вектора запроса с векторами страницы данных. В этом случае используются два регистра памяти по 32 разряда — регистр промежуточной памяти (РП1) и основной (РП2), а также узлы логической обработки.

Поясним работу алгоритма на примере поиска бóльших значений. Первоначально в РП1 запосятся все единицы, а в РП2 — все нули. В качестве слов опроса последовательно по разрядам предъявляются единицы (остальные разряды маскируются). В результате на каждом такте опроса из ФМАЗУ считываются одноименные разряды 32 векторов (a_{hi} для всех $h \in [1,32]$). В РП1 запоминается результат конъюнкции $C_{hi} = \bigcap_{j=1}^{i} G_{hj}$, где $G_{hj} = a_{hj}b_j + \bar{a}_{hj}\bar{b}_j$ — функция равнозначности j разрядов векторов страницы данных и вектора запроса. В РП2 суммируются результаты вычисления функции $d_{hj} = C_{h,j-1}a_{hj}\bar{b}_j$ (вычисляется функция $f_{hi} = \bigcup_{j=1}^{i} C_{h,j-1}a_{hj}\bar{b}_j$), т. е. на каждом такте добавляются номера векторов, совпавших с вектором запроса по предыдущим разрядам и содержащих бóльший i-й разряд (если таковые имеются). После 32 тактов опро-

са в РП1 остаются номера векторов, совпавших с вектором запроса по всем 32 разрядам, а в РП2 — номера строк с большими значениями векторов.

Экспериментальная модель базы данных строилась на основе 10^3 соединений. Из структурных формул этих соединений были выделены фрагменты с числом вершин от 2 до 7. В таблице представлены количество типов фрагментов в зависимости от их размера (числа вершин), средние размеры списков соединений и количество микроголограмм для их регистрации. ИНДЕКС ЗОНЫ размещался на магнитном диске и перед началом работы вводился в оперативную память ЭВМ СМ-4. ИНДЕКС УЧАСТКА, участки файла ФРАГМЕНТ и информация о соединениях регистрировались в ГН. При поиске в ГП были реализованы представленные выше методы сравнения данных по совпадению и в пределах «больше — меньше» непосредственно на ФМЛЗУ под управлением микроЭВМ. Она же выполняла операцию контроля ошибок, проводила анализ результатов сравнения и принятие решения.

Из кода предъявленного в запросе фрагмента последовательно выделялись по 2 байта, из которых формировались коды опроса страниц данных. В случае сравнения по совпадению результат опроса очередной страницы логически перемножался с результатом обработки предыдущих страниц, после чего проводился анализ на наличие совпавших кодов предъявленного и считанных из ГП фрагментов. При отсутствии совпадения процесс продолжался на следующей группе страниц и т. д., пока в файле ФРАГМЕНТ не обнаруживалась запись, содержащая ключ, идентичный коду предъявленного для поиска фрагмента.

После идентификации фрагмента считывались его регистрационный номер и список соединений, в состав структурных формул которых он входит. После этого из файла СТРУКТУРА считывались структурные формулы, названия, брутто-формулы и молекулярные веса выбранных соединений.

Экспериментально подтверждена возможность и эффективность работы ГП совместно с ОЭ-спецпроцессором, реализующим параллельную обработку страницы данных размером 32×32 бит. ЭВМ была разгружена от передачи и обработки большого объема данных, обработка в основном проводилась непосредственно в ГП, а в ЭВМ вводились результаты поиска. Время поиска соединений по фрагментам структурных формул определялось временем доступа в экспериментальном образце ГП и составило единицы секунд в пределах модуля памяти (информационная емкость ~2,5 Мбайта) и ~10 с в рамках всего объема ГП.

Повышение скорости поиска можно достичь путем применения более быстрых средств выборки данных. Так, например, при использовании электрооптического дефлектора время произвольного доступа в пределах модуля может составлять ~100 мкс [5], что позволяет увеличить скорость поиска примерно в 100 раз. Дальнейшее увеличение производительности системы достигается за счет использования более чувствительных парафазных фотоматриц, управляемых транспарантов с большим быстродействием и введения нескольких портов выборки данных из ГП.

Важно подчеркнуть, что время поиска данных для предложенного алгоритма слабо зависит от объема БД, в которой проводится поиск. Поэтому использование системы в конфигурации, показанной на рис. 1, тем эффективнее, чем больше объем обслуживаемой БД.

Авторы выражают благодарность участникам эксперимента А. А. Блоку, А. П. Литвинцевой, И. Б. Татарниковой.

СПИСОК ЛИТЕРАТУРЫ

1. **Петрова Е. М., Розенман М. И.** Автоматизированный банк данных STN INTERNATIONAL. — М.: ВНИИПАС, 1987.
2. **Гибин И. С., Кучерук Р. С., Потапов А. И.** Документальное голограммное запоминающее устройство: Сб. тр. 4-й Всесоюз. конф. по голографии.— Ереван: ВНИИРИ, 1982.— Т. 2.

3. **Блок А. А., Ванюшев Б. В., Гибин И. С. и др.** Испытания голографической системы архивной памяти емкостью 1 Гбайт // Тез. докл. VI Всесоюз. школы-семинара по оптической обработке информации.— Фрунзе: ФПИ, 1986.— Ч. 1.
4. **Коняев С. И.** Фотоматричное ассоциативное запоминающее устройство // Электрон. пром-сть.— 1988.— № 4.
5. **Грамматин А. П., Гусев В. К., Долгова Е. В. и др.** Голографическое запоминающее устройство с произвольным доступом к информации // ОМП.— 1988.— № 6.

Поступила в редакцию 15 мая 1989 г.

УДК 681.7.068

Е. С. АВДОШИН
(Тула)

ОПТОВОЛОКОННЫЙ ДАТЧИК ЗВУКА

Развитие оптических методов передачи и обработки информации ускорило разработку волоконно-оптических датчиков [1, 2] физических величин, и в частности световодных акустических датчиков [3—5].

Передача акустической информации по волоконным световодам позволяет достигнуть высокой чувствительности и помехозащищенности измеряемого сигнала. Световодные датчики можно использовать при сильных электромагнитных излучениях, в условиях взрыво- и пожароопасности. Дополнительно возникают возможности передачи больших объемов информации с применением мультиплексирования.

В данной статье описан волоконно-оптический датчик звука с вибрирующим световодом [6], конструкция которого показана на рис. 1. Датчик содержит металлический корпус *5*, в котором с помощью компаундного эпоксидного клея ВК-9 закреплены два световода *3* и *12*. Для передачи оптического излучения использованы многомодовые кварцевые световоды КВСП-50 с диаметрами сердцевины волокна 50 мкм и оболочки 125 мкм. Показатель преломления световода $n_c = 1{,}5$, а числовая апертура $NA = 0{,}2$. Каждый световод имеет длину 0,5 м. Зеркальные торцы на световодах выполнялись с помощью твердосплавного резца ВК-8 [7]. Качество сколов контролировалось микроинтерферометром МИИ-4 с увеличением ×500 и микроскопом МБС-9. На свободных концах световодов *3*, *12* установлены оптические соединители *2*, *13* типа ВОК — ВОК с внешним диаметром наконечника 2,5 мм для стыковки с источником света *1* и фотоприемником *15*.

Световод *3* закреплен в корпусе датчика в цилиндрическом капилляре *4* из нержавеющей стали. Внешний диаметр капилляра равен 2,5 мм, а внутренний точно соответствует размеру сердцевины волокна. После закрепления эпоксидом световода *3* внутри капиллярной трубки производилась шлифовка зеркального торца наконечника. На внешней поверхности капилляра выполнялась резьба, с помощью которой производились крепление световода *3* в корпусе датчика и регулировка зазора между световодами *3* и *12*.

Другой световод *12* закрепляется эпоксидом внутри датчика, образуя при этом консольную балочку *11* длиной 17 мм. Шток *10*, выпол-

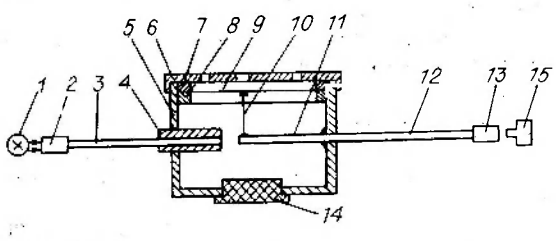

Рис. 1. Конструкция оптоволоконного акустического датчика:
1 — источник света; *2*, *13* — оптические соединители; *3*, *12* — световоды; *4* — капилляр, *5* — корпус датчика; *6* — крышка; *7*, *8* — кольца; *9* — мембрана; *10* — шток; *11* — балочка; *14* — заглушка; *15* — фотодиод