УДК 004.6

ФИЗИЧЕСКИЕ РЕСУРСЫ ИНФОРМАЦИОННЫХ ПРОЦЕССОВ И ТЕХНОЛОГИЙ

М.О. Колбанёв[a], Т.М.Татарникова[b]

[a] Санкт-Петербургский государственный экономический университет, Санкт-Петербург, 191023, Российская Федерация
[b] Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация, tm-tatarn@yandex.ru

Аннотация.

Предмет исследования. Рассмотрены базовые информационные технологии, автоматизирующие информационные процессы сохранения, распространения и обработки данных, с точки зрения требуемых им физических ресурсов. Показано, что изучение этих процессов с такими традиционными для современной информатики целями, как способность передавать знания, степень автоматизации, обеспечение информационной безопасности, кодирование, надежность и других, уже недостаточно. Причиной этого являются, с одной стороны, увеличение объемов и интенсивности информационного взаимодействия в ходе предметной деятельности людей и, с другой стороны, приближение к пределу производительности информационных систем, основанных на полупроводниковых технологиях. Актуальной проблемой современных инженерных разработок стало создание таких технических средств, которые не просто обеспечивают поддержку информационного взаимодействия, но и потребляют рациональные объемы физических ресурсов. Таким образом, объектом исследования являются базовые информационные технологии, обеспечивающие сохранение, распространение и обработку данных для поддержки информационного взаимодействия людей, а предметом исследования – физические временные, пространственные и энергетические ресурсы, необходимые для реализации этих технологий.

Используемые подходы. Предпринимается попытка за счет учета в явном виде объемов физических ресурсов, необходимых для изменения состояний носителей информации, расширить возможности традиционной методологии кибернетики, которая заменяет рассмотрение материальной составляющей информации перебором состояний информационных объектов.

Цель работы. Выработка общего подхода к сравнению и последующему выбору базовых информационных технологий сохранения, распространения и обработки данных с учетом не только требований к качеству информационного взаимодействия в определенной предметной области и степени использования технических средств, но и объемов потребляемых при этом физических ресурсов.

Основные результаты работы. Предложена классификация ресурсов, потребляемых базовыми информационными технологиями, по их физической природе на пространственные, временные и энергетические. Показано, что основными пространственными ресурсами применительно к базовым информационным технологиям являются плотность записи данных, распределение пользователей в зоне охвата и размер техпроцесса, временными – время гарантированного сохранения, время доставки данных и производительность обработчика, энергетическими – уровни энергетического барьера и сигнала и энергопотребление. Выделены ключевые физические ресурсы для базовых информационных технологий сохранения, распространения и обработки данных, к которым отнесены соответственно плотность записи, время доставки и энергопотребление. На примере технологий сохранения данных предложен подход к выбору такой информационной технологии, которая удовлетворяет требованиям пользователей к качеству информационного взаимодействия при рациональном потреблении физических ресурсов.

Практическая значимость. Результаты работы могут быть полезны специалистам, занимающимся проектированием и эксплуатацией высокопроизводительных систем вычислений, хранения и распространения данных; разработкой способов повышения эффективности существующих коммуникаций, включая мобильную и оптическую связь, методов и алгоритмов для сбора, хранения и интеллектуального анализа больших объемов данных; внедрением новых информационных технологий.

Ключевые слова: базовые информационные процессы, информационные технологии, сохранение данных, распространение данных, обработка данных, пространственные, временные и энергетические ресурсы информационных технологий, принцип Г. Мура, принцип Р. Ландауэра, точка Т. Стерлинга.

PHYSICAL RESOURCES OF INFORMATION PROCESSES AND TECHNOLOGIES

M.O. Kolbanev[a], T.M. Tatarnikova[b]

[a] Saint Petersburg State University of Economics, Saint Petersburg, 191023, Russian Federation
[b] Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, 190000, Russian Federation, tm-tatarn@yandex.ru

Subject of study. The paper describes basic information technologies for automating of information processes of data storage, distribution and processing in terms of required physical resources. It is shown that the study of these processes with such traditional objectives of modern computer science, as the ability to transfer knowledge, degree of automation, information security, coding, reliability, and others, is not enough. The reasons are: on the one hand, the increase in the volume and intensity of information exchange in the subject of human activity and, on the other hand, drawing near to the limit of information systems efficiency based on semiconductor technologies. Creation of such technologies, which not only provide support for information interaction, but also consume a rational amount of physical resources, has become an actual problem of modern engineering development. Thus, basic information technologies for storage, distribution and processing of information to support the interaction between people are the object of study, and physical temporal, spatial and energy resources required for implementation of these technologies are the subject of study.

Approaches. An attempt is made to enlarge the possibilities of traditional cybernetics methodology, which replaces the consideration of material information component by states search for information objects. It is done by taking explicitly into account the amount of physical resources required for changes in the states of information media.

Purpose of study. The paper deals with working out of a common approach to the comparison and subsequent selection of basic information technologies for storage, distribution and processing of data, taking into account not only the requirements for the quality of information exchange in particular subject area and the degree of technology application, but also the amounts of consumed physical resources.
Main findings. Classification of resources consumed by the basic information technologies is suggested according to their physical nature. They are: spatial, temporal and energy resources. It is shown that the main spatial resources for basic information technologies are: data recording density, the users' distribution in the coverage area and size of engineering process; temporal resources are: time of guaranteed saving, data delivery time and the handler efficiency; energy resources include: the barrier and the signal energy levels and power consumption. Key physical resources are highlighted for basic information technologies of data storage, distribution and processing that include, respectively, recording density, delivery time and power consumption. We suggest an approach to the selection of such information technology that meets the users' needs to the quality of information exchange with the rational consumption of natural resources. An example of data storage technology is given.
Practical relevance. The results can be useful for specialists involved in the design and operation of high-performance computing, storage and distribution of data, developing the ways of improvement for the effectiveness of existing communications, including mobile and optical communications, methods and algorithms for collecting, storing and smart analysis of large amounts of data, introduction of new information technologies.

Keywords: basic information processes, information technologies, data storage, data distribution, data processing, spatial, temporal and energy resources of information technologies, Moore's Law, R. Landauer's principle, T. Sterling's point.

Введение

Проблема снижения ресурсоемкости различных видов производств уже давно стоит перед всем миром. Для многих государств, регионов, отраслей промышленности и отдельных предприятий экономия ресурсов становится приоритетной задачей. Решение этой задачи в соответствии с Федеральным законом № 261-ФЗ об энергосбережении является обязанностью не только промышленности, но и муниципальных учреждений, государственных органов. Не является исключением и отрасль информационных технологий, где задача снижения ресурсоемкости стала приоритетной. Многие исследователи, изучая информацию, отмечали, что любые ее преобразования основаны на физических законах. Например, А.А. Ляпунов [1] указывал на ограничения пространства, времени и энергии при выполнении информационных технологий, поскольку невозможны концентрация слишком большой массы знаков в ограниченном объеме пространства, получение новых знаков и их передача в новый носитель за слишком маленькое время и регистрация новых знаков слишком маленькой энергией. Н.Н. Моисеев считал, что за исключением потребности изучения целенаправленных действий в живой природе и обществе можно обойтись без термина «информация» и протекающие процессы описывать с помощью законов физики и химии [2]. Р. Ландауэр [3] ставит знак равенства между информационными и физическими процессами, поскольку «информация физична». Определение понятия информации, которое следует из работ Н. Винера [4], явно связывает информацию с ее физическими свойствами: «Информация – это обозначение содержания, полученное нами из внешнего мира в процессе приспосабливания к нему нас и наших чувств».

Несмотря на такое понимание, физические свойства информации все же находились на втором плане исследований информатиков всю вторую половину XX века. Этому есть объяснение. Согласно закону Г. Мура, объемные характеристики информационных систем росли экспоненциально [5, 6]. Все возрастающие вычислительные возможности полупроводниковых технологий давали возможность рассматривать поведение кибернетических систем исключительно как нематериальное, а информацию как нематериальную субстанцию, которая, тем не менее, переводит системы из одного состояния в другое и самым существенным образом влияет на принятие решений.

Методология кибернетики основывается на трех базовых составляющих: системном подходе, прикладной математике и цифровых информационных технологиях. Такие прикладные математические теории, как теория информации, теория принятия решений, теория массового обслуживания, теория управления, моделирование систем, математическая и формальная логики, теории алгоритмов и автоматов, теории формальных языков и грамматик, социальная информатика, исследование операций и другие и сегодня составляют основу «информационного» образования. Общее у этих теорий – это «переборный» или «цифровой» метод:

1. сначала интеллектуал должен сформулировать цель исследования;
2. затем для достижения этой цели необходимо или выбрать некоторые состояния системы, или перебрать состояния, или упорядочить состояния, или исключить некоторые состояния, или синтезировать новые состояния, и т.п.

Весь смысл исследования прячется в цели, а вся «физика» – в умении сократить перебор, который для сложных систем является достаточно большим.

Цифровая информационная технология при таком подходе призвана методами прикладной математики реализовать алгоритмы перебора состояний системы, описанных цифровыми массивами данных.

Кибернетика учитывает смысловую составляющую информации только через цель, которая формулируется вне системы, а материальную оставляющую рассматривает как изменение состояний объектов, не связанное с их физической природой.

В последние годы стало очевидным существование некоторого предела возможностей полупроводниковых технологий, и это обстоятельство заставляет вернуться к физическим основаниям информационных преобразований. Главным системным ограничением для суперхранителей, суперпереносчиков и суперобработчиков данных нашего времени является энергопотребление. Уже сегодня крупные центры обработки данных, системы коммутации и маршрутизации, суперкомпьютеры в процессе своей работы потребляют десятки мегаватт электроэнергии. Один маршрутизатор операторского класса, например, каждый год потребляет столько энергии, сколько выделяется при сжигании десятков тонн угля.

Особенностью современных информационных технологий [7, 8], использующих принцип фон Неймана, является необходимость многократного сохранения, распространения и обработки данных. Это означает, что объемы энергии, потребляемые каждым информационным битом за время его жизненного цикла, увеличиваются многократно.

У многих информационных технологий можно проследить взаимную зависимость уровня энергопотребления с другими физическими ресурсами, описывающими пространственные и временные параметры систем [9].

Возможности преобразования информационных битов при их сохранении, распространении и обработке зависит сегодня не только и не столько от существования того или иного программного обеспечения, перебирающего состояния систем. Первостепенное значение приобретает наличие физических ресурсов, поскольку именно использование физических ресурсов обеспечивает перемещение данных как материальных объектов во времени, в пространстве и изменение формы представления данных [10, 11].

Рассмотрим с общих позиций те физические ресурсы, которые необходимы для выполнения функций базовыми информационными технологиями сохранения, распространения и обработки данных [12]. Очевидно, что физические ресурсы самым существенным образом влияют на технологические возможности реализации базовых информационных процессов [13]. Целью настоящей работы является выработка общего подхода к сравнению и последующему выбору базовых информационных технологий сохранения, распространения и обработки данных с учетом не только требований к качеству информационного взаимодействия в определенной предметной области и степени использования технических средств, но и объемов потребляемых при этом физических ресурсов.

Пространственные ресурсы

Любая информационная технология требует пространства. Пространственные ресурсы измеряются в единицах длины и расстояния, описывают способы размещения информационных объектов и должны эффективно использоваться при реализации информационных процессов. К числу основных пространственных ресурсов информационных технологий можно отнести следующие:

- для технологий сохранения – размер запоминающих устройств для записи и плотность записи данных;
- для технологий распространения – территория, в пределах которой организуется информационное взаимодействие пользователей (зона охвата) и распределение (плотность) пользователей на этой территории;
- для технологий обработки – размер техпроцесса и количество транзисторов, размещаемых в одном чипе.

Кроме того, любая технология характеризуется объемом технологических помещений и допустимой плотностью размещения в них оборудования.

Плотность записи данных – это количество бит, которое размещается на единице площади (или объема) запоминающего устройства (ЗУ). Очевидно, что плотность обратно пропорционально зависит от размера физического элемента, сохраняющего бит. Пока размер атома – это нижний теоретический предел увеличения плотности записи. Дальнейшее уменьшение размера 1 бита связано с расщеплением атома и переходом на квантовые технологии.

Для создания энергонезависимой памяти сегодня наиболее широко используются законы супермагнетизма. Плотность записи современного жесткого диска – менее 700 ГБ на дюйм^2. Максимальная теоретическая плотность в случае использования технологии HAMR (магнитной записи с подогревом) составляет 5–20 ТБ на дюйм^2 и может быть достигнута в скором будущем. Размер минимальной единицы хранения при этом должен быть порядка 10 нм. Уже реализована возможность сохранения данных в ячейке памяти, состоящей из 12 магнитных атомов, в то время как обычный жесткий диск для хранения одного бита данных использует сотни тысяч атомов [http://habrahabr.ru/post/136414/]. Повышение плотности позволит создавать компактные, быстрые и энергетически эффективные устройства. Полупроводниковая энергонезависимая память позволяет сократить энергозатраты при записи, хранении и чтении дан-

ных и обеспечивает плотность записи в зависимости от используемого техпроцесса. Техпроцесс 10 нм может быть освоен в ближайшие годы.

Зоной охвата современных информационных сетей является вся Земля. Это стало возможно благодаря созданию Интернет, в основе которого лежат:

- стандартизированные правила взаимодействия;
- единое адресное пространство;
- совместимость внутренних и внешних данных для всех сетей и добровольность объединения.

Отдельные сети имеют собственную зону охвата, управляются собственными администрациями и характеризуются используемыми информационными технологиями, составом пользователей, их количеством, интенсивностью взаимодействия, мобильностью, распределением по территории и др.

Применительно к телефонным сетям процедура связи вне зоны охвата «домашней» сети называется роумингом. Она требует предварительной взаимной договоренности между операторами, поскольку предполагает согласованное использование ресурсов нескольких сетей. Особый вид роуминга позволяет пользоваться мобильной связью на морском и воздушном транспорте, а также получить доступ к спутниковым сетям.

Размер техпроцесса определяет плотность транзисторов на одном кристалле. В соответствии с законом Г. Мура производительность кремниевых интегральных микросхем и количество транзисторов на одном кремниевом кристалле удваивается каждые 18 месяцев, а их стоимость при этом уменьшается на 50%. Рост количества транзисторов в одном чипе означает уменьшение и размеров единичного транзистора, и ширины контактных дорожек. Уровень техпроцесса 2011–2012 г.г. – это 22 нм, что соответствует размещению около 1,5 млрд транзисторов на 160 мм^2.

Уменьшение размера техпроцесса позволяет не только увеличивать плотность хранения данных на полупроводниках, но и создавать более сложные и эффективные архитектуры процессоров, в частности, имеющие несколько вычислительных ядер и уровней кэш-памяти. Кроме того уменьшение техпроцесса позволяет сократить энергопотребление за счет перехода на новые типы транзисторов, уменьшения напряжение питания, отключения в режиме бездействия отдельных ядер, кэш-памяти или участков интегрированного графического ядра и др.

Плотность размещения оборудования оценивается при помощи целой группы параметров. Это и количество вычислительных операций (вычислительная плотность), и объем потребляемой энергии (энергетическая плотность), и скорость информационных каналов (сетевая плотность) на единицу площади оборудования и др.

Рекорд вычислительной плотности 2013 г. – это 1 Пфлоп/с на одну стойку площадью 1 м^2. Стойка состоит из 1024 вычислительных узлов, имеет совокупную емкость локального файлового хранилища узлов 0,5 ПБ и обеспечивает отвод более 0,4 МВт тепловой мощности за счет использования прямого жидкостного охлаждения [http://www.rscgroup.ru]. Новой технологией, позволяющей сократить издержки на создание ИТ-инфраструктуры, являются модульные центры обработки данных. Производственное помещение строится из сэндвич-панелей, снабжается необходимым количеством серверов и прочего инфраструктурного оборудования и может располагаться в любом месте пространства при наличии доступа к сетевым и энергетическим мощностям [14].

В табл. 1 сведены ключевые физические и технологические пространственные ресурсы, характеризующие базовые информационные технологии.

Пространственные ресурсы	Базовые информационные технологии		
	Сохранение	Распространение	Обработка
Физические	Площадь (объем) ЗУ	Зона охвата	Размер техпроцесса
Технологические	Плотность (объем) записи	Плотность пользователей	Плотность транзисторов

Таблица 1. Пространственные ресурсы базовых информационных технологий

Временные ресурсы

Особенность времени как ресурса заключается в том, что его нельзя запасти впрок, оно расходуется непрерывно и необратимо. Управлять временем можно лишь планируя продолжительность тех или иных операций, в том числе с учетом случайных факторов.

Временные ресурсы информационных технологий – это время, необходимое для выполнения информационных процессов или их отдельных этапов и фаз. Эти ресурсы могут быть разделены на две группы, характеризующие, во-первых, время предоставления услуг (обслуживания) при сохранении, распространении и обработке данных и, во-вторых, время доступа к информационным услугам.

Для каждого из базовых информационных процессов время обслуживания имеет свое название, отражающее специфику процесса. Для сохранения – это время гарантированного сохранения, для распространения – время доставки данных, для обработки – производительность обработчика.

Время гарантированного сохранения – это период времени, который начинается в момент записи данных на ЗУ и продолжается до тех пор, пока данные могут быть найдены на ЗУ, считаны и интерпретированы пользователем. Это время зависит от времени «жизни» минимальных единиц хранения, т.е. времени, в течение которого они сохраняют установленное состояние.

Примерами современных долговечных хранителей данных являются диски типа M-Disc, которые записывают данные на слое минерального материала, подобного камню, и гарантируют сохранность файлов на протяжении 1000 лет [http://www.mdisc.com].

Еще более выносливым является стеклянный диск. Он не имеет минерального слоя, устойчив к природным катастрофам, пожарам и излучениям, выдерживает условия открытого космоса, температуры, близкие к абсолютному нулю, и излучение Солнца. Одна из компаний дает гарантию в 100 лет на накопители, созданные на базе флэш-памяти с антикоррозийной защитой. Электроны в плавающем затворе транзисторов сохраняются тем дольше, чем ниже температура хранения [15]. Согласно новому открытию можно синтезировать частицу ДНК и записать в нее экзабайты данных. Затем в лиофилизированной форме ДНК можно сохранять теоретически тысячи лет.

Время доставки данных – это период времени, который начинается в момент поступления сигнала в канал связи и заканчивается по достижению данными заданной точки пространства (адресата).

Время доставки по сети связи включает время передачи данных от источника информации в канал связи, время перемещения сигнала по каналу между сетевыми центрами и время управления движением сигнала в сетевых центрах, таких как маршрутизаторы, серверы или телефонные станции. И в электрических, и в оптических сетях собственно время перемещения сигнала по каналу связи равно скорости света. Задержки передачи сигналов связаны с необходимостью обрабатывать адресную и другую управляющую информацию, сопровождающую данные при использовании коммутируемых сетей [16].

Пропускная способность канала – это наибольшая скорость передачи данных, измеряемая в бит/с, т.е. количество данных, которые сеть может передать за единицу времени между двумя оконечными устройствами. Она достигается при использовании оптимальных для данного канала настроек источника информации, когда на каждом такте работы канала каждый символ переносит максимально возможное количество бит данных.

Производительность (показатель, обратный времени обработки данных) – это количество операций обработки в секунду.

Основной задачей процесса обработки данных является получение нового массива данных из исходного при помощи некоторых алгоритмов. Для решения этой задачи в архитектуре фон Неймана задействованы вычислительные элементы и память, объединенные коммутационной сетью (интерконнектом). Вычислительные элементы – это процессоры, каждый из которых содержит несколько вычислительных ядер, память – это иерархически организованная система хранения программ и данных, включающая регистры, кэши, основную и внешнюю памяти [17].

В сложной архитектуре компьютеров скорость счета зависит не столько от свойств элементной базы, сколько от способов объединения процессоров, памяти и интерконнекта. Это обстоятельство подтверждают данные, приведенные в табл. 2.

Год	1997	2011	Изменение
Техпроцесс	250 нм	22 нм	↓ в 10 раз
Тактовая частота процессоров	1 ГГц	1–4 ГГц	↑ в 2,5 раза
Время переключения транзисторов	$250 \cdot 10^{-15}$ с	$3 \cdot 10^{-15}$ с	↓ в 100 раз
Максимальная производительность суперкомпьютера	1 Тфлоп/с	33 Пфлоп/с	↑ в 33000 раз

Таблица 2. Динамика изменения вычислительных характеристик компьютера

Дальнейшее увеличение производительности суперкомпьютеров требует увеличения числа вычислительных узлов и развития методов программирования параллельных вычислений, однако главным резервом увеличения производительности является уменьшение времени доступа при обращении к памяти и к интерконнекту для перемещения данных между вычислительными узлами.

Появление суперкомпьютеров производительностью до 1 Эфлоп/с (10^{18} флоп/с) ожидается до 2020 г. Обсуждается возможность приближения суперкомпьютеров к зетта-масштабу (10^{21} флоп/с) до 2030 г.

Время доступа – это интервал времени между моментами поступления заявки на предоставление информационной услуги до момента начала ее реализации. Оно зависит от способа использования ресурсов информационных технологий, таких как объем запоминающих устройств, каналов и процессоров или энергии [18]. Если за некоторым пользователем заранее закреплен достаточный физический и технологический ресурс, то время доступа будет малой величиной, которой можно пренебречь. Однако, как прави-

ло, информационные системы организуют доступ многих пользователей к ограниченному количеству ресурсов. При этом возникают коллизии, и пользователи вынуждены ожидать освобождения нужных им ресурсов, если они уже используются другими пользователями. Если количество ресурсов системы рассчитано таким образом, что время доступа не превышает согласованной с пользователем величины, то систему называют системой реального времени [19].

В табл. 3 сведены ключевые физические и технологические временные ресурсы, характеризующие базовые информационные технологии.

Временные ресурсы	Базовые информационные технологии		
	Сохранение	Распространение	Обработка
Физические	Время «жизни» минимальной единицы хранения	Время доставки знака	Время переключения транзисторов
Технологические	Гарантированное время хранения	Пропускная способность	Производительность

Таблица 3. Временные ресурсы базовых информационных технологий

Энергетические ресурсы

Вслед за увеличением объема и интенсивности информационных потоков и охватываемой ими территории малая энергия, требуемая для управления малыми информационными потоками, перерастает в большую. В результате информационные системы потребляют сегодня колоссальное количество электроэнергии.

По всему миру на снабжение информационного и телекоммуникационного оборудования сейчас тратится около 160 ГВт, что составляет 8% от всей вырабатываемой электроэнергии, и эти показатели продолжают быстро расти [20]. По различным оценкам, к 2020 г. потребность оборудования информационных систем в электроэнергии увеличится более чем в два раза и достигнет 400 ГВт. Основными потребителями являются оконечные устройства, центры обработки данных и оборудование сетей.

Бит как единица оценки количества данных уже недостаточен для сравнения возможностей и эффективности информационных систем. Имеют значение и физический размер бита, и время его гарантированного сохранения, и энергия, необходимая для сохранения, передачи и обработки бита.

Эффективность информационных систем связана сегодня с фактическим потреблением ими физических ресурсов (в первую очередь электроэнергии**)** и оценивается, например:

- объемом энергии, потребляемой в расчете на единицу информационных услуг;
- стоимостью транзакций в киловатт-часах или объеме выбросов углерода;
- объемами выбросов углерода в пересчете на один сервер или на группу пользователей;
- соотношением энергопотребления информационного оборудования и инженерных систем, поддерживающим его работу;
- энергопотреблением на 1 м^2 площади технических помещений и т.д.

Р. Ландауэр в 1961 г. показал [7], что расход энергии в процессе вычислений связан с уничтожением битов данных, и сформулировал следующий принцип: «Независимо от физики и технологии вычислительного процесса при потере 1 бита данных в процессе вычисления как минимум выделяется энергия, равная $k_{\text{В}}T\ln 2$, Дж», где $k_{\text{В}}$ – постоянная Больцмана, определяющая связь между температурой и энергией (порядка $1{,}3807\cdot10^{-23}$ Дж/К); T – температура, при которой ведутся вычисления (300 К = 26,85°С). Остальные операции (копирование, установка, перенос и др.) требуют сколь угодно мало энергии при достаточно малой скорости протекания. В 2012 г. были представлены результаты экспериментов, подтвердивших этот результат.

Объяснение этого принципа очевидно. Обработка бита является операцией над двумя битами. Поскольку биты материальны и имеют размер, то и для перехода в состояние 0 или 1 они должны получить энергию. При выполнении операции один из входных битов превращается в результат операции на выходе, а другой теряется, и его энергия выделяется в виде тепла и излучений.

Количество тепла, которое выделяется при стирании 1 бита, очень мало. Но в архитектуре фон Неймана значения битов в памяти переписывается с огромной частотой, и выделяемую энергию уже нельзя не учитывать. Тем более, что уровень энергетических затрат на обработку одного бита при технологии 22 нм лежит в пределах $\left(k_{\text{В}}\cdot T\cdot 10^{5}\right)-\left(k_{\text{В}}\cdot T\cdot 10^{6}\right)$, т.е. в миллион раз больше, чем минимально возможный, а с учетом сопутствующих потерь на 10 Пфлоп/с сегодня тратится порядка 10 МВт электроэнергии.

То, что на языке программистов называется изменением данных в памяти, на физическом уровне означает рассеивание в пространстве тепла и излучений. Таким образом, энергопотребление – это главное системное ограничение для будущих информационных технологий.

При сохранении минимальные единицы хранения данных должны быть отделены друг от друга и от среды достаточно сильными энергетическими барьерами. Вероятность их искажения зависит от многих физических и химических факторов и определяется законом Аррениуса. Считается, что для сохранения минимальных единиц хранения в течение миллиона лет нужен энергетический барьер 60–70 КВт.

При распространении данных сигналы подвергаются воздействию помех. Для достоверной доставки следует или увеличивать уровень сигнала на этапе физического кодирования символов сообщений, или применять алгоритмы помехоустойчивого кодирования, фактически заменяя один символ группой символов. И в одном, и в другом случае требуется энергия. Теплотехнический консорциум (сообщество инженеров-теплотехников, занятых в области производства компьютерной техники) исследовал технические характеристики компьютерного оборудования и проанализировал тенденции развития новых средств вычислительной техники. Результаты исследований показали, что наибольшую удельную тепловую нагрузку создает телекоммуникационное оборудование.

Обработка данных. Т. Стерлинг в 2009 г. на конференции по суперкомпьютерам в Гамбурге предположил, что экзафлопсный рубеж окажется пределом развития современных суперкомпьютеров. Точка Стерлинга – это условное ограничение производительности суперкомпьютера, построенного на доступных сегодня технологиях [20]. Это ограничение следует из принципа Р. Ландауэра.

Пусть задано энергопотребление суперкомпьютера в 20 МВт = $20 \cdot 10^6$ Вт, что близко к пределу при электронных технологиях. Если разделить эту величину на $150\, k_{\text{в}} T$, где 150 – это эмпирический коэффициент минимального уровня энергии на обработку одного бита для надежной работы компьютера, то при комнатной температуре получим около $4 \cdot 10^{26}$ операций/с. Выполнение одной операции над 64-разрядным числом с плавающей точкой требует 20 тыс. однобитовых операций, поэтому в пересчете на операции с плавающей точкой получаем $2 \cdot 10^{22}$ флоп/с. Последнюю оценку надо уменьшить более чем на 2 порядка, учитывая недостатки материалов и технологий производства. В результате приходим к выводу Т. Стерлинга: максимальная производительность суперкомпьютера при современных технологиях находиться в пределах 32–128 Эфлоп/с и, вероятно, никогда не превзойдет величины 64 Эфлоп/с. Это означает, что закон Г. Мура перестанет действовать в близком будущем.

В табл. 4 сведены ключевые физические и технологические энергетические ресурсы, характеризующие базовые информационные технологии.

Энергетические ресурсы	Базовые информационные технологии		
	Сохранение	Распространение	Обработка
Физические	Энергетический барьер	Уровень сигнала	Энергозатраты на обработку бита
Технологические	Энергопотребление при сохранении	Энергопотребление при распространении	Энергопотребление при обработке

Таблица 4. Энергетические ресурсы базовых информационных технологий

Ресурсная модель базовых информационных технологий

Для описания ресурсного обеспечения базовых информационных технологий может быть использован параллелепипед (рис. 1), грани которого отображают нижние и верхние границы пространства, времени и энергии, необходимые информационным технологиям на некотором этапе их развития. К соответствующим значениям следует стремиться при выборе информационной техники [19]. В этой модели ось ординат (S) отображает пространственные, ось аппликат (F) – энергетические, ось абсцисс (T) – временные параметры технологии. Точки, лежащие в пределах объема параллелепипеда и имеющие координаты $0 \le S \le S^{\max}$, $0 \le T \le T^{\max}$, $0 \le F \le F^{\max}$, соответствуют некоторым уже реализованным или еще разрабатываемым технологиям.

В качестве примера рассмотрим технологии сохранения больших данных, которым также требуются все большие объемы физических ресурсов. При этом необходимо учитывать требования, в том числе к:

- плотности записи, поскольку от нее зависят размеры и количество носителей;
- величине гарантированного времени сохранения данных;
- величине энергозатрат, которые необходимы для записи/считывания данных на носитель и для защиты носителя от внешних воздействий между моментами записи и считывания.

Перечисленные пространственные, временные и энергетические параметры цифровых технологий сохранения зависят друг от друга. Улучшение любого из них может быть, как правило, достигнуто только за счет ухудшения других. Это хорошо демонстрируют, например, технологии полупроводниковой памяти SLC, MLC, TLC, в которых увеличение плотности записи достигается за счет снижения гарантированного времени хранения. То же самое можно отнести и к магнитной памяти, где увеличение плотности (уменьшение площади доменов) ведет к появлению взаимного влияния магнитных полей и, следовательно, снижает гарантированное время хранения и увеличивает энергетические затраты для защиты (восста-

новления) данных в процессе хранения. Полупроводниковая память характеризуется малым энергопотреблением по сравнению с магнитной, но, заменяя магнитную память полупроводниковой, следует учитывать, что за экономию энергии придется заплатить меньшим временем гарантированного хранения. Уменьшение энергопотребления дисковыми массивами возможно за счет остановки или уменьшения скорости вращения дисков, но это увеличивает время доступа к данным, и т.д.

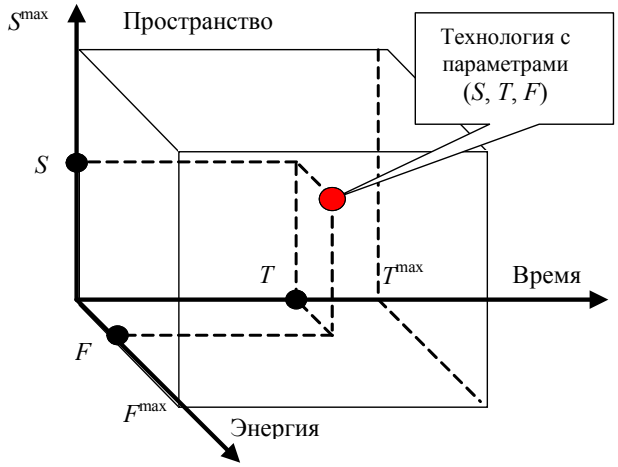

Рис. 1. Модель физических ресурсов информационных процессов

В общем случае эффективность технологии сохранения тем выше, чем больше плотность и гарантированное время хранения данных и меньше затрачиваемая при этом энергия, поэтому требованиям пользователей к реализации процесса сохранения соответствует параллелепипед с гранями S^{cox}, T^{cox}, F^{cox} на рис. 2.

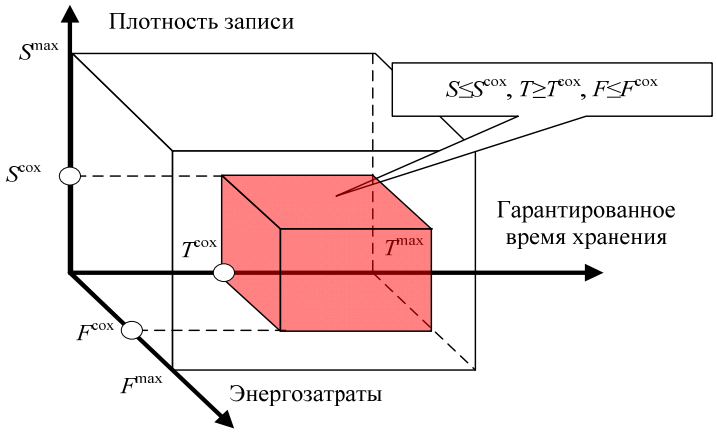

Рис. 2. Допустимая область параметров для технологии сохранения данных

В качестве базовых технологий для долговременного сохранения данных могут использоваться магнитные диски и ленты, полупроводниковая память, компакт и М-диски и др. В каждом из этих технологических сегментов существует множество альтернативных реализаций, в том числе и у разных производителей. Проведенные исследования показали, что каждая технология характеризуется собственными физическими параметрами:

- магнитные диски обеспечивают высокую плотность, но малое время хранения;
- магнитные ленты гарантируют длительное время хранения, но характеризуются и сравнительно значительным временем доступа;
- полупроводниковая память позволяет экономить энергию, но не способна на долговременное хранение;
- компакт-диски гарантируют длительное время хранения, но имеют ограниченную плотность записи;
- М-диски наиболее долговечны, но энергозатратны в процессе записи и т.д.

Примерами технических решений, принятых и реализованных при сохранении больших данных, могут служить следующие системы.

Большой адронный коллайдер является источником огромного количества научных данных о результатах столкновений элементарных частиц. Оцифрованные данные только о тех столкновениях, которые дали интересные с точки зрения физики результаты, поступают со скоростью около 50 событий в

секунду. В главном центре хранения и обработки эти данные записываются на магнитную ленту. К настоящему времени на десятках тысяч картриджей уже собрано более 100 ПБ данных. Доступ к картриджам автоматизирован: они хранятся в специальных подвальных помещениях на полках, откуда их достает робот. Энергопотребление системы хранения составляет 3,5 МВт. Фрагменты данных копируются на перекрывающий кэш диска для доступа и распределения между исследовательскими центрами по всему миру. Особенности данной системы – это отсутствие жестких ограничений на время обработки данных, что позволяет уменьшить энергозатраты.

Другой пример – это система долговременного хранения данных ColdStorage компании Facebook. Она располагается в отдельном здании и оптимизирована с точки зрения уменьшения энергопотребления и увеличения плотности размещения данных, а не производительности и доступности. Для этого используются магнитные диски, которые не рассчитаны на постоянную эксплуатацию, но позволяют менять скорость вращения, увеличивать количество дисков в одной стойке и уменьшать количество одновременно вращающихся дисков. В результате выбора именно такой технологии хранения для энергопитания массива дисков емкостью 1 ЭБ требуется примерно 0,375 МВт вместо 1,5 МВт.

Еще одно решение от Facebook – это экспериментальное хранилище, которое состоит из трехсот тысяч оптических дисков, где хранятся 30 ПБ данных. Нужный диск с требуемыми файлами находит робот. В перспективе система оптического хранения сможет сохранять до 150 ПБ. Эта система заметно увеличивает время доступа к затребованным файлам, но позволяет увеличить гарантированное время сохранения и на 80% снизить энергопотребление.

Выбор того или иного решения должен основываться на использовании обобщенного критерия или сравнении доступных технологий друг с другом по всем физическим характеристикам, которые могут быть отображены в виде параллелепипеда, подобного представленному на рис. 2. В процессе такого сравнения следует, используя принцип Парето, удалить из рассматриваемого набора такие технологии, которые заведомо хуже других по пространственным, временным и энергетическим характеристикам в совокупности. Если технологий, превосходящих другие хотя бы по одному показателю, останется достаточно много, то следует применить один из методов многокритериального выбора.

Набор параметров задачи оптимизации зависит при этом от вида базовой технологии сохранения данных. Оптимизируемыми параметрами могут являться, например, скорость вращения и количество одновременно вращающихся дисков, плотность дисков на стойке, пространственные, временные и энергетические характеристики робототизированных систем, объемы сохраняемых данных и др. В качестве критериев эффективности можно выбирать потребляемые физические ресурсы системы, а в качестве ограничений – вероятностно-временные характеристики производительности и доступа, размеры производственных помещений, возможности силовых агрегатов и т.п.

Предложенный подход развивает традиционные схемы оптимизации производительности информационных систем за счет выбора числа обслуживающих устройств, их быстродействия и надежности без учета объемов потребляемых при этом физических ресурсов и применим к широкому кругу базовых информационных процессов и технологий.

Заключение

Выявленный тренд развития базовых информационных технологий показывает, что учет затрат на обеспечение информационных технологий физическими ресурсами становится существенным при проектировании мощных информационных систем, таких как, например, системы хранения данных, центры обработки данных или суперкомпьютеры. Их эффективность зависит не только от возможностей информационных технологий в части объемов хранения, скорости передачи и обработки данных, но и от объемов занимаемого пространства, временных параметров физических процессов и потребляемой энергии. Исходя из этого, при проектировании подобных информационных систем следует:

- рассматривать весь спектр доступных информационных технологий сохранения, распространения и обработки;
- учитывать в процессе выбора физические параметры технологий;
- согласовывать физические параметры технологий с допустимыми параметрами задач пользователя;
- учитывать взаимную зависимость пространственных, временных и энергетических характеристик технологий.

Литература

1. Ляпунов А.А**.** Проблемы теоретической и прикладной кибернетики. М.: Наука, 1980. 335 с.
2. Моисеев Н.Н. Человек и ноосфера. М.: Молодая гвардия, 1990. 351 с.
3. Landauer R. Information is physical // Physics Today. 1991. V. 44. N 5. P. 23–29.
4. Винер Н. Кибернетика, или управление и связь в животном и машине. М.: Советское радио, 1958. 215 с.
5. Moore G.E. Cramming more components onto integrated circuits // Proceedings of the IEEE. 1998. V. 86. N 1. P. 82–85.

6. Kish L.B. Moore's law and the energy requirement of computing versus performance // IEE Proceedings: Circuits, Devices and Systems. 2004. V. 151. N 2. P. 190–194.
7. Landauer R. Irreversibility and heat generation in the computing process // IBM Journal of Research and Development. 2000. V. 44. N 1. P. 261–269.
8. Советов Б.Я., Колбанёв М.О., Татарникова Т.М. Технологии инфокоммуникации и их роль в обеспечении информационной безопасности // Геополитика и безопасность. 2014. № 1 (25). С. 69–77.
9. Bean J., Dunlap K. Energy-efficient data centers: a close-coupled row solution // ASHRAE Journal. 2008. V. 50. N 10. P. 34-36+38+40-42.
10. Schmidt R., Beaty D., Dietrich J. Increasing energy efficiency in data centers // ASHRAE Journal. 2007. V. 49. N 12. P. 18–21+24.
11. Gea-Banacloche J., Kish L.B. Future directions in electronic computing and information processing // Proceedings of the IEEE. 2005. V. 93. N 10. P. 1858–1863.
12. Советов Б.Я., Колбанёв М.О., Татарникова Т.М. Модель физических характеристик сигналов // Материалы VIII Санкт-Петербургской межрегиональной конференции «Информационная безопасность регионов России (ИБРР-2013)». Санкт-Петербург, 2013. С. 64–65.
13. Kish L.B., Granqvist C.G. Does information have mass? // Proceedings of the IEEE. 2013. V. 101. N 9. P. 1895–1899.
14. Belady C.L., Beaty D. Roadmap for datacom cooling // ASHRAE Journal. 2005. V. 47. N 12. P. 52–55.
15. Тысячелетний накопитель. Новейшие разработки в области хранения информации // Chip. 2012. № 6. С. 114–119.
16. Колбанёв М.О., Татарникова Т.М., Воробьёв А.И. Модель обработки клиентских запросов // Телекоммуникации. 2013. № 9. С. 42–47.
17. Tatarnikova T., Kolbanev M. Statement of a task corporate information networks interface centers structural synthesis // IEEE EUROCON 2009. 2009. Art. 5167903. P. 1883–1887.
18. Советов Б.Я., Колбанёв М.О., Татарникова Т.М. Оценка вероятности эрланговского старения информации // Информационно-управляющие системы. 2013. № 6 (67). С. 25–28.
19. Богатырев В.А., Богатырев А.В. Функциональная надежность систем реального времени // Научно-технический вестник информационных технологий, механики и оптики. 2013. № 4 (86). С. 150–151.
20. Лёвшин И. Многоточие Стерлинга // Суперкомпьютеры. 2010. № 3 (15). С. 6–8.

Колбанёв Михаил Олегович – доктор технических наук, профессор, профессор, Санкт-Петербургский государственный экономический университет, Санкт-Петербург, 191023, Российская Федерация, mokolbanez@mail.ru

Татарникова Татьяна Михайловна – доктор технических наук, доцент, профессор, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация, tm-tatarn@yandex.ru

Mikhail O. Kolbanev – D.Sc., Professor, Professor, Saint Petersburg State University of Economics, Saint Petersburg, 191023, Russian Federation, mokolbanez@mail.ru

Tatiana M. Tatarnikova – D.Sc., Associate professor, Professor, Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, 190000, Russian Federation, tm-tatarn@yandex.ru

Принято к печати 11.05.14
Accepted 11.05.14