Appl Phys A (2011) 102: 865–875
DOI 10.1007/s00339-011-6297-0

Applied Physics A
Materials Science & Processing

Artificial cognitive memory—changing from density driven to functionality driven

L.P. Shi · K.J. Yi · K. Ramanathan · R. Zhao · N. Ning · D. Ding · T.C. Chong

Received: 26 October 2010 / Accepted: 22 December 2010 / Published online: 5 February 2011

Abstract Increasing density based on bit size reduction is currently a main driving force for the development of data storage technologies. However, it is expected that all of the current available storage technologies might approach their physical limits in around 15 to 20 years due to miniaturization. To further advance the storage technologies, it is required to explore a new development trend that is different from density driven. One possible direction is to derive insights from biological counterparts. Unlike physical memories that have a single function of data storage, human memory is versatile. It contributes to functions of data storage, information processing, and most importantly, cognitive functions such as adaptation, learning, perception, knowledge generation, etc. In this paper, a brief review of current data storage technologies are presented, followed by discussions of future storage technology development trend. We expect that the driving force will evolve from density to functionality, and new memory modules associated with additional functions other than only data storage will appear. As an initial step toward building a future generation memory technology, we propose Artificial Cognitive Memory (ACM), a memory based intelligent system. We also present the characteristics of ACM, new technologies that can be used to develop ACM components such as bioinspired element cells (silicon, memristor, phase change, etc.), and possible methodologies to construct a biologically inspired hierarchical system.

L.P. Shi (✉) · K.J. Yi · K. Ramanathan · R. Zhao · N. Ning · D. Ding · T.C. Chong
Data Storage Institute, A*STAR (Agency for Science, Technology and Research), Singapore 117608, Singapore
e-mail: SHI_Luping@dsi.a-star.edu.sg

1 Introduction

Memory technology is one of the most important pillars of the IT industry and has been widely used in consumer electronic products. The current mainstream memory technologies include optical storage, magnetic storage, and solid state memory [1–3]. In practice, almost all computing systems use a variety of memory types—organized in a storage hierarchy—as a trade-off between performance and cost. Generally, the lower a storage is in the hierarchy, the lesser its bandwidth and the higher its density. As such, solid state memories are historically used as internal memories due to their high speed and relatively high cost, while optical and magnetic storage devices are typically used as external storage devices due to low cost and slow speed. Up until now, increasing the data density has been the long-term trend in the history of memory development.

Optical data storage has been established for about 20 years, primarily as a read only data storage. It can provide Read Only Memory, Write Once Read Many and Re-Writable functions [4, 5]. Due to the unique features of large capacity, long life-time, portability, low cost, and noncontact data retrieval, optical disks are widely used in multimedia to store digitized audio, video, and images. It dominates the market for distributing prerecorded audio and video. The normal method to increase the data density of optical disks is to utilize shorter wavelength laser diode and higher numerical aperture objective lens to reduce the beam spot size. The third generation Blu-Ray optical disks use a wavelength of 405 nm and numerical aperture of 0.85 to achieve 25 GB/side [6]. However, it is not practical to further increase the data density by using shorter wavelength (i.e., UV laser diode) because it involves a significant change in the optical system which has not been developed yet. Currently, three possible approaches are under development to further the

data density increase. The first approach is near-field optical recording that uses a near-field optical fiber probe or solid immersion lens (SIL) to achieve higher density and speed [7, 8]. The most critical challenge is the servo system for controlling the position of SIL optical pick-up and the gap between SIL and the surface of media. The second approach is multilayer recording [9]. Recording up to sixteen layers has been demonstrated. However, although it is possible to achieve above 500 GB/disk by combining near field technology and multi-layer technology, many challenges, such as yield, need to be overcome. The third approach is holographic optical recording that can provide very large storage density and high speed [10]. Information is recorded in the holographic medium through the interference of two coherent light beams. The most promising holographic recording technology is to use photopolymer. However, until now only write once holographic technology can meet the application requirements, which seriously limits its applications.

Hard disk drive (HDD) has been the primary type of magnetic data storage devices. It is the first choice of secondary storage device in computer systems due to high performance, low cost, nonvolatile, and durable characteristics. Its areal density has been increased by around 300 million times from the first HDD product in 1956. Currently, the HDD is of areal density at 600 Gb/in^2 in commercial product and at 1.2 Tb/in^2 in lab Spinstand demonstration. Researchers are moving toward 10 Tb/in^2 areal densities and beyond, which needs the bit size of around 64 nm^2 and below [11]. However, the magnetic recording technology is facing its physical limitation, the superparamagnetism, starting from the areal density at a few 10s Gb/in^2. With the innovation of emerging technologies, such as perpendicular recording, energy assisted magnetic recording, and bit patterned medium, the technology solutions toward increasing write ability and reducing grain number per bit have pushed the recording density to current level and it is expected to realize 10 Tb/in^2 in 8 to 10 years time [11–14]. Moving forward, there are many challenges ahead at 10 Tb/in^2. The energy assisted writer has to create less than 20 nm of writing bubble with the effective field strength more than 50 KOe. If the bit patterned medium has to be used, the low cost and durable nanopatterning technology with the etching width not more than 4 nm has to be achieved. Also, the head media spacing has to be 2 to 3 nm with mechanic spacing budget less than 0.5 nm [11].

Although volatile memories (dynamic/static random access memory as typical examples) are still important solid state memories, recently the demand for non-volatile memory (NVM) has been greatly increased due to the rapid growing market for portable electronic products, such as digital cameras and cellular phones. Flash memory is the dominating NVM technology. However, it is believed that it will soon run into fundamental technological barriers when the technology node reaches about 20 nm. As such, much effort has been put to develop alternative NVMs such as Magnetic Random Access Memory (MRAM), Ferroelectric Random Access Memory (FeRAM) and Phase Change Random Access Memory (PCRAM), Organic Thin-Film Memory, Molecular Memory, etc. [15–19] High density and high scalability are the most important criteria to determine which memory technology may succeed as the replacement. Among the alternative memories, PCRAM is considered as one of the best candidates due to its near-ideal memory advantages: fast access time, low power, low cost, high endurance, high scalability, and good data retention [20]. The most significant advantage is its high scalability. Since the energy required for phase transformation decreases with the volume of phase change materials, the write current scales with the device size that facilitates memory scaling. Considering the significant material property change of the phase change materials at the nanoscale [21, 22], the scaling limitation of PCRAM is estimated to be 5 nm. 128 Mb PCRAM chips have been commercialized recently. The latest fundamental discovery on spin transfer effect has put spin torque transfer MRAM (STTMRAM) as another potential next generation NVM [23, 24]. It has advantages of small device size, low writing current density, fast writing speed, and good scalability. In recent years, remarkable progress in STT switching with MgO magnetic tunnel junctions and increasing interest in STTMRAM has been witnessed [25]. It has been demonstrated that with the reduction of device size the current needed to switch the memory device could be reduced. If only consider from the material point of view, it can be estimated that STTMRAM could be scaled down to about 5 nm.

2 Artificial cognitive memory to go beyond density driven

As discussed above, for the past several decades, the focus of research in memory technologies has been on increasing the data density. Numerous methods have been pursued to minimize the physical size in order to increase the areal density. However, although researchers have broadened their efforts beyond scaling and process innovation to include novel cell structures, it is expected that existing data storage technologies may reach their physical limits in around 15 to 20 years due to miniaturization. To continue the technology advancement, it is required to develop new data storage architecture and explore a new development trend that is different from density driven.

2.1 Drawing inspirations from the brain for future memory technologies

In fact, nature may have already provided a solution. One of the most advanced data storage mechanisms known to us

is the human memory. Human memory has been viewed to be the basis for human intelligence, involving various functions, such as storage, perception, learning, and association. To appreciate the relevance of human memory to our problem, a brief comparison of current memory technologies and the human memory needs to be performed.

The sole function of physical memory is the storage of binary data, such that it can be transferred in space and time. Human memory, on the other hand, is versatile, performing the functions of storage and processing, as well as higher level cognitive capabilities, such as perception, learning, prediction, and association. The information is stored in a non-binary, structure-based, adaptive, and experience dependent manner and is retrieved using contextual cues. An amodal, invariant representation format is employed, which stores information in a parallel, distributed, and hierarchical fashion.

Human memory can also be expressed in different forms, such as episodic, semantic, procedural, and sensory memory. Each of these has a short and long term component, which can be loosely correlated to the volatile and non-volatile aspects of physical memory. Table 1 briefly summarizes the similarities and differences between human memory and physical memory.

The multifunctional nature of human brain and its capability to operate with low power consumption inspires this research for a new trend in memory research. This technology, called Artificial Cognitive Memory (ACM), is intended to emulate the functional nature of human memory.

Table 1 A brief comparison of human memory and physical memory

	Human memory	Physical memory
Typical usage	Information storage and processing	Information storage
Storage Medium	Short term (working)/long term (semantic, procedural, episodic, declarative, etc.) memory	Volatile/NVM
Operation method	Digital/analog hybrid (Spikes/synapses)	Digital
Memory format	Non-binary	Binary
Adaptability	Yes	No
Learning	Yes	No
Data access methods	Parallel	Serial/Parallel
Physical layout	Categorical/hierarchical	Array/Matrix
Data access time/ms	10 [26]	~ 0.01 [26]
Power consumption	Low	High
Functions	Cognition, information storage, processing, perception, prediction, learning, etc.	Information storage

2.2 Proposed characteristics of ACM

For our human memory, incoming signals are not only stored, but also processed and associated with the existing content of the system. With this concept in mind, we outline the proposed characteristics of the ACM.

1. ACM is a memory-based intelligent system that can store and process information concurrently.

Unlike that in the current physical memory, the information in ACM is stored in a way that can facilitate to achieve cognitive functions, such as invariant information retrieval.

2. ACM is adaptive.

Human memory has high adaptive capabilities. ACM should be able to adapt to its own needs and those of its environment.

3. ACM system can grow dynamically.

Neurons in human brain grow and self organize dynamically during different periods of life. The complexity of ACM should be able to grow dynamically as well, so that it can be best suited in various circumstances, i.e., a system built from a few cells to billions of cells.

4. ACM has learning and prediction capability.

Learning from experiences and predicting the future events are the two of key features of intelligence. To be intelligent, ACM should have these capabilities.

5. ACM can interact and associate within and between old and new multimodal data.

Different compartments of the brain are structurally associated by means of neuronal links, enabling the brain to associatively process multiple modalities of information. In addition to that, incoming information is also associated with prior data by means of changed synaptic strengths. Interactions between new and old information, such as association and decision making after these interactions should be well addressed by ACM. For example, the state of the ACM at time t, $S(t)$, is dependent on interactions with prior states.

$$S(t) = f(S(t-1), S(t-2), \ldots, S(t-k)), \quad k \to \infty \quad (1)$$

6. ACM should have a hierarchical structure with distributed information.

The consensus in the neuroscience and psychological communities is that the brain processes information in a par-

allel, distributed, and hierarchical fashion. The ACM should implement this distributed processing capability.

7. ACM is capable to tolerate errors.

The distributed nature of the brain, coupled with its provision for redundancy, contributes to the robustness in the brain. Observing patients with semantic dementia tells us that the brain loses information in a graceful manner [27], losing the general details of the information later than specific ones. The ACM should be robust and have capability to tolerate the system errors.

8. ACM unifies software and hardware.

The brain works as an interplay between neural circuits (hardware) and different learning mechanisms (software). ACM embodies this harmonic integration of hardware and software through a convergence of bottom up and top down approaches.

9. ACM should have low power consumption.

The power consumption is much lower for human brain than that of physical memory. For example, although the supercomputer consumes power as high as megawatts, it still could not reach the intelligence level of human brain which only consumes 20 watts [28]. This significant difference is another driving force for us to develop ACM. Therefore, ACM should be built to have low power consumption with a scale of several to tens of watts or even less.

10. New evaluation metrics for ACM need to be developed.

Clearly, the traditional evaluation metrics, such as density, speed, power consumption, etc. will be insufficient for evaluating ACM, although they must be considered from the beginning. Besides conventional metrics, some novel evaluation criteria need to be developed.

Table 2 shows the summary of the ACM properties and compares them between the current physical memory memory and ACM.

Thus, it is possible that the future driving force for memory technology development might evolve from density driven to functionality driven. ACM is likely to be this kind of future memory.

Table 2 Comparison between current physical memory and ACM

	Current memory	ACM
Objective	Data storage	Memory based intelligence
Function	Storage	Storage and processing
Adaptability	Static	Dynamic with learning capability
Prediction capability	Nil	Multiple construction /reconstruction
Physical layout	Array/Matrix	Parallel, distributed, and hierarchical
Association capability	Nil	Associative interactions between and within old and new information
Data access	Linear/Limited parallel/Non-associative	Parallel, Associative
Power consumption	High	Low
Robustness to error	Low	High
Evaluation metrics	Density, Access speed	To be developed

3 Development of ACM

Current artificial intelligence and neural network techniques are based on computing. ACM, instead of being a solely computing solution, will be a memory-based, brain inspired intelligent system with the integration of hardware and software.

The organization of human nerve system (Fig. 1), can be viewed from a macro level perspective, i.e., the scale of the nerve system (1 m), the molecular perspective (1 Å), or from the subsystem (10 cm), neuron assembly (1 mm), neuron (100 μm), and synapses (1 μm) point of view [29]. Modules at increasingly macro levels are composed of modules at micro levels, with each module performing specific or combinational functions. The construction of ACM should be based on the same principle, with materials in the bottom, overall system on the top and individual modules in between.

In order to develop ACM, we need to identify and classify the functional modules of the neocortex, build bioinspired element cells to emulate selected neural functions, connect these element memory cells with adaptive synapses, and design a hierarchical architecture to form a system. The system should have information storage, information processing and cognitive functions. The most challenging of these is the implementation of cognitive functions, as they are not well understood even in the fields of neuroscience and psychology. Therefore, the development of the ACM should, at this stage, be based on current understanding of the human brain, and must evolve along with new discoveries in neuroscience and psychology. ACM could be developed in a strategic method, starting from basic element cells, to simple system with plastic networking, to complete the hierarchical system.

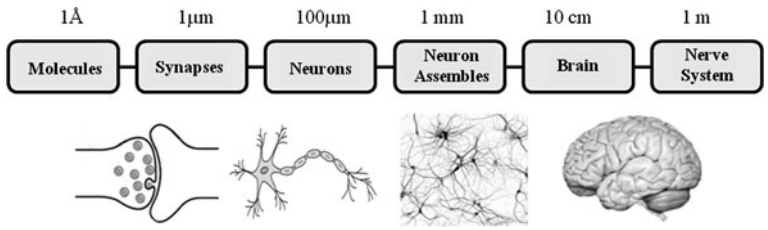

Fig. 1 Organization of human nerve system

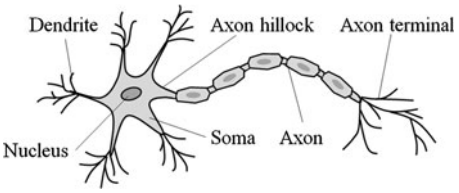

Fig. 2 Structure of a biological neuron

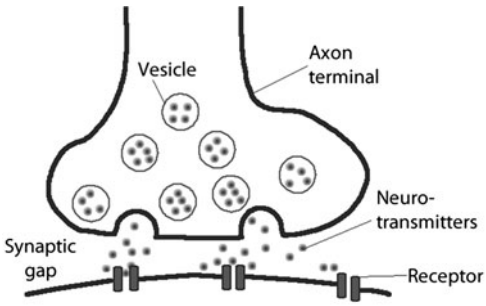

Fig. 3 Structure of a biological synapse

3.1 Element cells

Developing bioinspired element cells with neuron like properties is the first step to building the complex cognitive memory system that the ACM represents.

3.1.1 Biological background

Two fundamental units, neuron and synapse, in the human brain play essential roles in learning and in the formation of memory [30]. Neurons, as the core component of the nervous system, are electrically excitable cells and are able to respond to stimuli such as sound and light, to conduct impulses, and to communicate with each other via synapses. Synapses are membrane-to-membrane junctions that allow rapid transmission of electrical and chemical signals.

A typical neuron possesses a soma (the cell body which contains the cell nucleus), dendrites and an axon, as shown in Fig. 2. Dendrites are filaments that arise from the cell body, giving rise to a complex "dendritic tree," which are responsible for receiving signals from other neurons. An axon is a special cellular filament that arises from the cell body at a site called the axon hillock, where the input signals are integrated.

Every neuron maintains a voltage gradient across its membrane, attributable to the metabolically-driven differences in ion concentrations of sodium, potassium, chloride, and calcium within the cell. At rest, the cell membrane is polarized maintaining a negative interior charge of -70 mV. If a stimulus causes sodium ion channels within the membrane to open, leading to a flow of sodium ions into the cell, the potential of the neuron becomes more positive, and thus the membrane is depolarized. When enough depolarization accumulates to bring the membrane potential up to threshold, action potentials are triggered. Action potentials propagate as waves along axons, which cause chemicals called neurotransmitters to be released at synapses (Fig. 3). A cell that receives a synaptic signal may be excited, inhibited, or otherwise modulated.

It is believed that memory is formed by the modification of synapses and neuronal circuits through plasticity [30]. Synaptic plasticity can be expressed as the strengthening and weakening of synapses. Synaptic strength is affected by the amount and type of neurotransmitter in the synaptic gap. It is also influenced by the sensitivity of the post synaptic receptors.

3.1.2 Modeling

Neurons are highly nonlinear devices that communicate with each other based on action potentials: the firing of a neuron alters the dynamics of its connected neurons through synapses, either excitatory or inhibitory. High level cognitive functions, mental states, and information processing are coded by firing dynamics of a large group of neurons. Concerning memory, it is likely that experience physically modifies the strength of synapses, or in other words, the connectivity of the corresponding neural ensemble. A data reading procedure (recall) could be spontaneous firings of neurons according to a configured connectivity by a prior experience. In this case, the subject is considered to have a strong impression about the previous experience. It is also possible that a recall is triggered by external stimuli. For example, we see a person with a familiar face, but cannot remember the name. After several seconds of thinking, the name suddenly comes out of the mind. This scenario can be modeled as follows: a previous experience indeed modifies the connectivity of a group of neurons. However, the modifications are not strong enough to trigger spontaneous firing. Therefore, we cannot remember the person's name immediately.

A variety of patterns are input to the group of neurons as efforts to remember the name, including verbally pronouncing different family names, and to remember when or where this person is seen, etc. When one set of input patterns become in resonance with the weakly configured neuronal connectivity, the name comes out of the mind. Along with the name, it is common experience that many details of previous meetings with the person also come out.

Action potential that is a consequence of neurons' firing plays a central role in modeling the dynamics of neurons. There are roughly two kinds of models for action potential [31–39]. One of them more focuses on channel dynamics and biophysics basis, for example, the Hodgkin–Huxley model and compartment model. A second kind of model more focuses on interneuron connection, convergent, and divergent network, and threshold crossing detection, for example, leaky integrated and firing (LIF) model, integrated information, and artificial neural network. These models have played important roles in recent debates about the origin and responsible variability in cortical neurons [40–44]. Other models of spiking neurons [45] include, but not limited to: Izhikevich [46], Wilson [47], Hindmarsh-Rose [48], Morris-Lecar [49], FitzHugh–Nagumo [50], and Resonate-and-Fire [51] models.

ACM intends to mimic neuronal dynamics from at least two perspectives: 1. developing nonlinearity and integration using new materials and devices; and 2. synthesizing neuronal connectivity through wiring and switching.

3.1.3 Physical implementation

Several approaches, represented by silicon, memristor, and phase change technologies, have been used to build devices to emulate some biological behaviors of neuron and synapse. These technologies are introduced in this section.

Silicon The operation method between metal oxide semiconductor field effect transistors (MOSFET) in subthreshold range based on silicon technology is analogous to ion channel properties of neurons, which has sparked intensive interests in scientific and engineering communities to build analog silicon neuron devices [52–54]. The first analog silicon neuron, developed in 1991 was able to emulate efficiently the ion currents that cause nerve impulses and control the dynamics of their discharge [54]. More powerful analog neuron devices based on silicon technology have subsequently been proposed [52, 53, 55, 56].

In general, at least four transistors are required for one artificial analog neuron and at least two transistors for one artificial synapse [57]. In the human brain, there are 10^{11} neurons and 10^{15} synapses working together. Implementing this high number of 10^{15} transistors on a single chip is a big challenge [58]. Even if this can be implemented, a serious consideration is the heavy power consumption, and the heat dissipation problem will become detrimental.

Artificial neurons and synapses can also be built digitally when transistors operate in cut-off and saturated regions. This type of digital neurons can be implemented by a computer or field programmable gate array (FPGA) devices through neural modeling [59, 60].

Memristor Since its discovery by the R.S. William group at Hewlett Packard, the memristor [61] is believed to be able to spark a wave of technological innovations. The memristor has the capability to remember the charge or flux it experiences [61–64], which has been used to emulate synaptic plasticity. The memristor is therefore a good candidate to emulate certain functions of neuron behavior.

The sizes of memristor cells/devices are primarily at nanoscale, allowing them to be integrated with extremely high density. For example, recently, an integrated density of $10^{12}/\text{cm}^2$ has been achieved using CMOS in combination with molecule-scale two-terminal memristors (CMOL) [63]. This scale will be close to the level of a large scale neural ensemble.

Metal dioxides and polymers can be used to build memristor devices. Very recently, memristors made from magnetic materials, called spintronic memristors, were also proposed [65, 66].

Employing CMOS neurons in combination with memristor synapses can yield a system that mimics spiking timing dependent plasticity (STDP) [64, 67, 68]. Other than artificial synapses, memristors can also be used to perform logic operations [69].

Phase change materials The properties of chalcogenide-based phase change materials have been widely investigated in the last few decades [70–74] in the development of optical disks and PCRAM. It has also been proposed to build cognitive devices—Ovonic Cognitive Devices—based on phase change materials as early as in 2004 [75]. The idea of this cognitive device originated from the behavior similarity between biological neuron and synapse and phase change materials: accumulation, firing and weighing. These behavior similarities allow us to use only phase change materials, perhaps with a limited fraction of other auxiliary elements to build cognitive devices.

Functions of phase change cognitive device consist of non-binary storage, non-binary information processing, encryption, addition, division, modular arithmetic, factoring in parallel and adaptive learning capability [76]. Recently, phase change cognitive devices have been demonstrated to realize nearly all functions of memristors, such as implicit logic operations [77].

Different phase change materials, or phase change materials with artificial structures [78, 79], enable the possibility

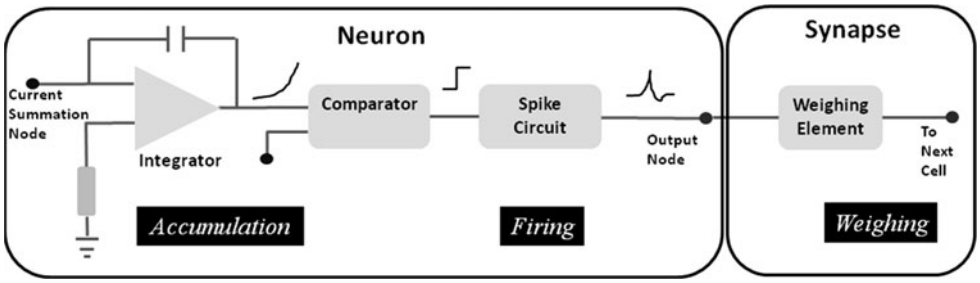

Fig. 4 Schematic diagram of one example of an element cell

to develop the devices emulating the behavior of different molecular properties of neurons, such as threshold voltage, response time to stimulus, structure, connection, neurotransmitters, etc.

Phase change material-based cognitive devices are expected to have high integrated density, good read/write cycles, long retention time, and low power consumption.

Other materials Besides the above mentioned approaches, neurons and synapses can also be behaviorally emulated from other materials. Ag_2S atomic switches were used to build a nanodevice that can mimic short and long term memory [80]. To emulate the dynamic functions such as excitatory postsynaptic current, inhibitory postsynaptic current or STDP with small dimension, low power and inexpensive devices, a synaptic transistor based on ionic/electronic hybrid materials was designed and fabricated by integrating a layer of ionic conductor and a layer of ion-doped conjugated polymer, onto the gate of a silicon-based transistor [81].

3.1.4 Bio-inspired element memory cells in the ACM

Element memory cells, ensembles of neuron-like computing nodes—artificial neurons and synapses—should be the core component of ACM. The element memory cell is not a reproduction of a neuron or synapse, but a minimal configuration of components required for memory storage and processing. By integrating many element memory cells into a hierarchical system with adaptive connections, an ACM device could come into being. To ensure that the element cell has a good scalability, a communication interface with which the element cell could interact with other cells or modules and an operation core with which the information can be processed, interpreted and stored should be well defined.

An element memory cell should be implemented by materials that are suitable and will be the basic element to build more complicated modules. In principle, various materials of silicon, metal dioxide, polymer, and phase change or new materials can be used to fabricate element memory cells.

The element memory cell consists of the key behavioral properties of the neuron—excitatory/inhibitory inputs, an integrator, a threshold comparator, an action potential generator, a firing compartment, a timing synchronizer and outputs. It also contains the key behavioral property of a synapse—neurotransmitter governed weighing.

The element cell should contain at least three functions of information communication (in the form of inputs and outputs), processing (integration and firing), and storage (as synaptic weights) with a compact structure and adaptive capabilities (dynamically changing synaptic weights), operating in asynchronous or synchronous manner.

Figure 4 shows the schematic diagram of one example of an element cell. The operation of this cell is as follows: input signals from the presynaptic cells are accumulated by the integrator. Upon reaching the threshold voltage, the digitized signal is generated by the comparator, and processed by the spiking circuit (spiking signal generator). The spikes are output to the synapse (or weighing element), whose weight can be changed dynamically. The output signals from the weighing element goes to the next stage for further processing and storage.

It is desirable that one material could perform all the functions, eliminating the complicated fabrication processes. Memristors and phase change materials might be the suitable candidates to build the cells.

3.2 Hierarchical system

While much research effort has gone into both the cellular modeling of neurons and synapses, as well as physical realizations of these, it is unclear how neurons can be connected together to achieve cognitive functions [82, 83].

System level cognitive models, using machine learning techniques, have generally taken a behavioral approach and have performed well on tasks such as object categorization. However, to truly understand the role of neurons in cognition, it is necessary to step out of the behavior modeling approach taken by the machine learning community and step into understanding how networks of neurons represent information such that they can perform complex tasks.

One possible way to obtain this understanding is to characterize the cognitive memory system embedded in the neocortex and hippocampus of the brain by using mathematical and simulation tools. Obeying neurobiological and psychological mechanisms underlying the memory system, we can

build up network models and synapse based algorithms to perform high level cognitive tasks such as associative memory and learning of spatial and temporal patterns at the modular level. The modeling, based on low-level neural circuits, has the property of scalability and feasibility. Brain-based models and algorithms will help us to prototype cortical microcircuits for incorporating into ACM.

In this section, we cite some literatures on information organization in the cortex, and provide some insights on how to physically link element cells into a cortical architecture to achieve cognitive systems.

3.2.1 Biological background

Even though the number of neurons in the brain is very large, only a very small fraction of neurons are activated in response to a cognitive task. Studies show, for instance, that stimulating a certain behavior repeatedly activates the same neurons. This observation lays the foundation for neural assemblies and the ideas of cortical microcircuit theory [84, 85].

Why are ensembles of neurons capable of encoding sensory stimuli so effectively, despite the fact that their inputs come from a large number of weak synapses? Research in Neuroscience and Psychology suggest that reliable representation can be attributed to representations and categorizations based on multiple similarities [86–88]. Although this may lead to seemingly redundant layouts of circuitry [89], the redundancy also results in a certain robustness of the memory, as can be viewed in patients with dementia [27, 90]. The concepts of multiple representations and neural assemblies provide a powerful framework for neural information processing for the development of brain inspired memory architectures.

Cognitive functions come from spiking neurons working together in asynchronous and synchronous manners, both in spatial and temporal domains. Recent investigations on neuronal consciousness derived the integrated information theory [91], which attempts to explain how groups of neurons can be wired together to generate consciousness [92, 93]. State-of-the-art technologies on neural recording, neural stimulation and neural signal processing techniques will also allow us to design advanced tools to identify cognitive functions behind various spiking characteristics of small groups of in-vivo neurons [94–97].

Two important discoveries in neuroscience have shed light on the structure of the cortex—the part of the brain that is associated with memory. They are the theory of the cortical column and the Hubel Weisel model of hierarchical information organization. Neuroscientist Vernon Mountcastle showed that the cortex is made up of replicating structures—the cortical column, which is the fundamental unit of information processing [98]. In 1965, David Hubel and Torsten Weisel found that "simple S cell" neurons in the cat's cortex showed specific firing preference according to the angle of the line presented. "Complex C cell" neurons responded to lines moving in one direction, showing how the visual system builds an image from simple features to complex objects. The combination of S cells and C cells, whose signals propagate up the hierarchy allows for scale and position invariant object recognition [99]. This work provided insight on how information propagates in the hierarchical structure of the cortex and has been the inspiration for the Hubel Weisel models of hierarchical memory.

3.2.2 System modeling

The mammalian cortex is a fundamentally modular and hierarchical structure. Modeling the modular (cortical column), and the hierarchical (system level) properties of the cortex is a good direction to pursue for ACM devices. As such, devices that replicate properties of the element cell, discussed in the previous section, can be connected to develop a brain inspired memory system.

Many literatures can be found on developing cortex-inspired learning models for visual pattern recognition applications [100–102]. These models are modular and make use of the S and C cell hierarchy described by Hubel and Weisel. They, however, focus mostly in the domain of pattern recognition. Preliminary results [103, 104] show that the Hubel Weisel models can be used out of the domain of vision, making them suitable candidates for the kind of multi modal and integrative representation that the brain performs. In the implementation for these models, it should be noted however, that cognition consists of aspects beyond pattern recognition. As ACM is not intended to be solely a pattern or image recognition engine, but a device for storage and processing, part of the system level modeling work explores how higher level cognitive tasks can be incorporated into the hierarchical cortical architecture.

Furthermore, the cortical models are driven at the modular level by machine learning algorithms such as connectionist, Bayesian networks, radial basis functions, and support vector machines [100–104]. ACM design, however, is bolder—it aims to have at its base neuron- synapse connections—implemented in hardware. Integrating the modular-system level architectures to neural correlates therefore forms a significant step in ACM research.

3.2.3 Physical implementation

Classical computer memories rely on the static approach and are very different from human memories. The ACM circuits should be somewhat reminiscent of human memory, covering aspects of human memory such as the ability to encode temporal-spatial data, semantic memory, episodic memory, etc.

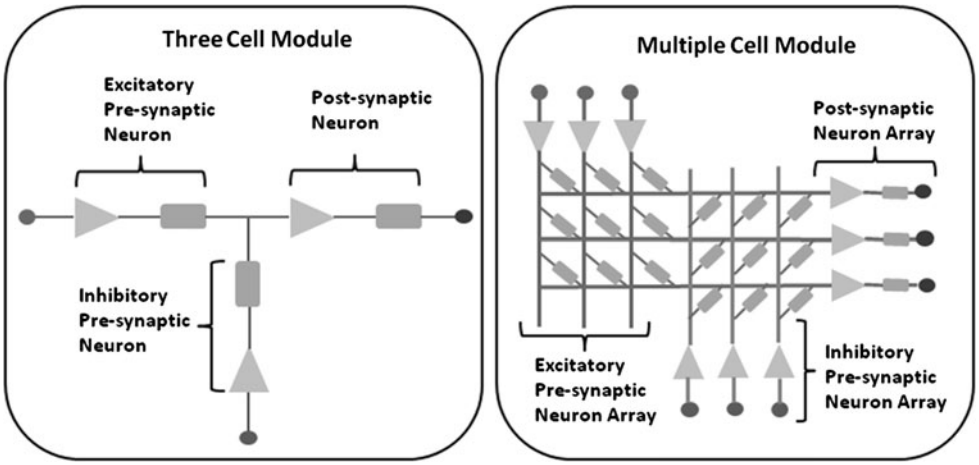

Fig. 5 Schematic diagram of three cell and multicell modules

Rather than using Boolean logic, precise digital representations, and clocked operations, as with computers, human memory carries out robust and reliable computation using hybrid analog/digital unreliable components. Human memory also emphasizes distributed, event-driven, collective, and massively parallel mechanisms, and makes extensive use of adaptation, self organization, and learning. These characteristics of memory are believed to be crucial for cognitive functions.

To implement a module and system level characterization of the ACM, individual element cells are connected and grouped to a certain scale to from circuits to achieve a certain function such as association or attention. Figure 5 shows an example of neuron modules.

Different functions should be realized by different neural circuits, but not from neurons with different structure. This is inspired from the fact that, in human memory, neurons used for different functions are fundamentally similar but their connections and circuitry are different. How to achieve the cell- structure-independent of functional building blocks is a daunting task.

Some modules/groups should be connected together, with others are separated, either physically and non-physically, to realize subsystems and ultimately form a complex ACM system. Figure 6 shows the schematic diagram of such a system where multiple element cells combine to form a module and multiple modules form the hierarchy. It should be noted that Fig. 6 shows a 7-layer structure as a statement that we are not constraining ourselves to the six layered architecture of the neocortex.

One possible way toward high performance ACM is through CMOL: a hybrid of integrated molecular device, nanowires and CMOS. A typical CMOL circuit includes an advanced CMOS subsystem with two, mutually perpendicular, arrays of parallel nanowires, and similar nanodevices formed at each crosspoint of the nanowires. CMOL's high density has the capability of simulating large scale neuronal

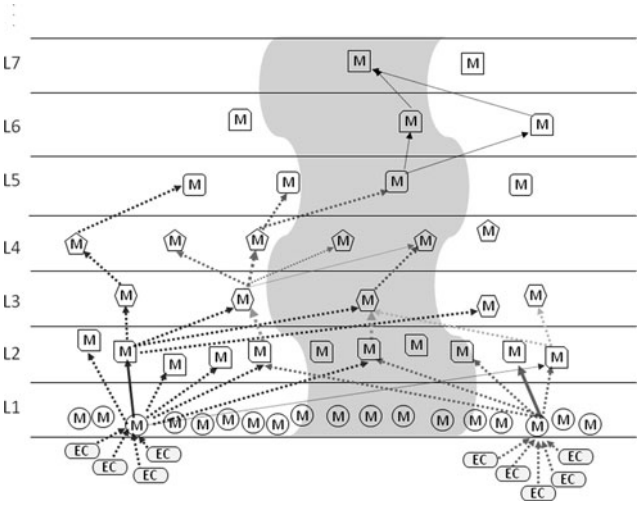

Fig. 6 Hierarchical feedforward architecture as an example of the ACM system design, emulating the layers of the cortex. The *shaded area* represents a cortical column. EC: Element cell, M: module

networks. This can be a platform for incorporating neural intelligence through non linear networks which require a large number of neurons.

In summary, the concrete steps of physically building ACM will include: (1) materials, (2) cells, (3) circuit, (4) subsystem, (5) system, and (6) evaluation environments. In parallel, investigation steps for theoretical computation include modeling the following: (1) material, (2) neuron, (3) circuit, and (4) neocortex. The algorithms based on these models should be able to be implemented and verified by off-the-shelf electronic components, such as FPGA and application specific integrated circuits.

It should also be considered that the brain is an evolving, self organizing organ that grows constantly from prenatal to adulthood. Similarly, at its highest, ACM should be an evolving and dynamic system—with the capability to grow, specialize, self-organize, and differentiate into differ-

ent neural circuits to achieve increasingly higher level cognitive capabilities.

4 Discussion

Realizing intelligence is important for human evolution. Throughout time, human lives have become better with the use of tools. The more intelligent the tools, the greater the degrees of freedom humans possess to move towards better health, knowledge, productivity, and safety. Electricity has extended the physical capability of humans including their energy and functionality. Computers have extended human brain capability in terms of processing speed and memory scale. Networking has enlarged the connectivity of human brains. ACM will extend human brain capability in terms of intelligence. Emulating the human brain has been the dream of computer science researchers.

Today's computers can perform billions of operations per second, but they are still no match for even a young child when it comes to tasks such as pattern recognition or visual processing. The human brain is also millions of times more energy-efficient and far more compact than a typical personal computer. Compact, efficient electronics based on the brain's neural system (namely neuromorphic microchips) could yield implantable electronic microchips to restore some brain's functions, like vision, hearing, and memory impairment or loss, as well as other smart sensors. Undoubtedly, developing ACM will be an important step toward this objective.

From another point of view, modeling and studying cognitive memory system will also extend and expand human's capability to understand his own brain intelligence. Most of current discoveries about brain functions were made from experiments by electroencephalography (EEG) and functional Magnetic Resonance Imaging (fMRI) analysis, which face inevitable constraints. Simulating the artificial cognitive memory will facilitate the overcoming of such constraints, thus allow our ability to explore human brain intelligence leap forward to an unprecedented level.

Upon beginning, some interesting questions wait for answers:

- The link between neurons and human intelligence has not been understood yet. Without this knowledge, what are the possible and practical approaches to build ACM?
- What mechanisms of encoding and decoding information used in the brain can be adopted and incorporated into ACM?
- What basic and necessary bioinspired element cells do we need to build? Are there alternatives to bio inspired cells for ACM functions?
- How can the ACM reflect the distributed storage and information processing nature of the brain?
- What determines the permanency level of information in the ACM?
- Memory in the brain has several aspects—episodic, semantic, etc. How much of this should be reflected in the ACM and how to design them?
- What cognitive functions that we really desire and should prioritize in order to develop ACM?
- Forget function is substantial in human memory for refreshment and attention, should it be incorporated into ACM?

In summary, the current memory technologies have achieved great success in terms of density and speed by reducing feature size per bit. ACM with cognitive functions might be a future memory extended from the current high density memory technologies. It could be built based on bioinspired element cells to emulate neuron functions, plastically connect with adaptive synapses to form a hierarchical architecture to achieve cognitive functions. It will also be developed to evolve along with the new discoveries in neuroscience and psychology. ACM as a memory based cognitive system may pave a new way to realize cognitive intelligence.

Acknowledgement This work was supported by Agency for Science, Technology, and Research (A*Star), Singapore (Grant No: 0921570130). The authors would like to thank Prof. Yang Zhi at the Department of Electrical & Computer Engineering of National University of Singapore, Dr. Yuan Zhimin, and Dr. Liu Bo for their valuable discussions.

References

1. Information storage industry consortium (INSIC) Optical Data Storage Roadmap (2006)
2. Z.Z. Bandic, R.H. Victora, Proc. IEEE **96**(11), 1749 (2008)
3. *International technology roadmap for semiconductor* (2009)
4. *CD Standard* (Rainbow Books)
5. DVD standard
6. Blu-ray Disc, Basic Format Specification version 1.0 (2002)
7. B.D. Terris, H.J. Marnin, G.S. Kino, Appl. Phys. Lett. **65**, 388 (2002)
8. E. Betzig, J.K. Trautman, Science **257**, 189 (1992)
9. A. Mitsumori et al., Jpn. J. Appl. Phys. **48**, 03A055 (2009)
10. J.F. Heanue, M.L. Bashaw, L. Hesselink, Science **265**, 749 (2009)
11. Z.M. Yuan et al., IEEE Trans. Magn. **45**(11), 5038–5043 (2009)
12. M.H. Kryder et al., Proc. IEEE **96**(11), 1810–1835 (2008)
13. H.J. Richter et al., IEEE Trans. Magn. **42**(10), 2255–2260 (2006)
14. R. Wood et al., J. Magn. Magn. Mater. **235**(1–3), 1–9 (2001)
15. C.P. Collier et al., Science **285**, 391 (1999)
16. G.W. Burr et al., J. Vac. Sci. Technol. B **28**, 223 (2010)
17. J.F. Scott, J. Phys., Condens. Matter **18**, R361 (2006)
18. J.Y. Ouyang et al., Nat. Mater. **3**, 918 (2004)
19. W.J. Gallagher, S.S.P. Parkin, IBM J. Res. Dev. **50**, 5 (2006)
20. R. Bez, IEDM Tech. Dig. (2009)
21. L.P. Shi, T.C. Chong, J. of Nanoscience and. Nanotechnology **7**, 65 (2007)
22. S. Raoux, J. Jordan-Sweet, A. Kellock, J. Appl. Phys. **103**, 114310 (2008)

23. J.G. Zhu, Proc. IEEE **96**, 1786 (2008)
24. Y. Huai et al., Appl. Phys. Lett. **84**, 3118 (2004)
25. E. Chen et al., IEEE Trans. Magn. **46**, 1873 (2010)
26. T. Bilski, *Digital and Biological Storage Systems—A Quantitative Comparison*. (Bioetics, 2007)
27. T.T. Rogers, J.L. McClelland, Nat. Rev., Neurosci. **4**, 310 (2003)
28. R.J. Douglas, K.A. Martin, Curr. Biol. **17**, R496 (2007)
29. P.S. Churchland, T.J. Sejnowski, *The Computational Brain* (MIT Press, Cambridge, 1992)
30. E.R. Kandel, Science **294**, 1030 (2001)
31. A.L. Hodgkin, A.F. Huxley, J. Physiol. **117**, 500 (1952)
32. B.W. Knight, J. Gen. Physiol. **59**, 734 (1972)
33. G.S. Oxford, J. Gen. Physiol. **77**, 1 (1981)
34. H.C. Tuckwell, *Introduction to Theoretical Neurobiology* (Cambridge University Press, Cambridge, 1988)
35. K. Nagy, J. Membr. Biol. **96**, 251 (1987)
36. L. Lapicque, J. Physiol. Pathol. Gen. **9**, 620 (1907)
37. M.A. Wilson, J.M. Bower, The simulation of large-scale networks, in *Methods in Neuronal Modeling*, ed. by C. Koch, I. Segev (MIT Press, Cambridge, 1989), p. 291
38. R.D. Keynes, F. Elinder, Proc. Biol. Sci. **265**, 1393 (1998)
39. S. Michalek et al., Eur. Biophys. J. **28**, 605 (1999)
40. G. Bugmann, C. Christodoulou, J.G. Taylor, Neural Comput. **9**, 985 (1997)
41. L.F. Abbott, Brain Res. Bull. **50**, 303 (1999)
42. M.N. Shadlen, W.T. Newsome, J. Neurosci. **18**, 3870 (1998)
43. T.W. Troyer, K.D. Miller, Neural Comput. **9**, 971 (1997)
44. W.P. Softky, C. Koch, Neural Comput. **4**, 643 (1992)
45. W. Gerstner, W. Kistler, *Spiking Neuron Models* (Cambridge University Press, Cambridge, 2002)
46. E.M. Izhikevich, IEEE Trans. Neural Netw. **14**(6), 1569–1572 (2004)
47. H.R. Wilson, J. Theor. Biol. **200**(4), 375–388 (1999)
48. R.M. Rose, J.L. Hindmarsh, Proc. R. Soc. Lond. B, Biol. Sci. **237**(1288), 267–288 (1989)
49. C. Morris, H. Lecar, Biophys. J. **35**(1), 193–213 (1981)
50. R. Fitzhugh, Biophys. J. **1**(6), 445–466 (1961)
51. E.M. Izhikevich, Neural Netw. **14**(6–7), 883–894 (2001)
52. C. Rasche, R. Douglas, Analog Integr. Circuits Signal Process. **23**, 227 (2000)
53. E. Farquhar, P. Hasler, IEEE Trans. Circuits Syst. **52**, 477 (2005)
54. M. Mahowald, R. Douglas, Nature **354**, 515 (1991)
55. J.H.B. Wijekoon, P. Dudek, Neural Netw. **21**, 524 (2008)
56. R. Douglas, M. Mahowald, C. Mead, Annu. Rev. Neurosci. **18**, 255 (1995)
57. C. Bartolozzi, G. Indiveri, Neural Comput. **19**, 2581 (2007)
58. R.W. Williams, K. Herrup, Annu. Rev. Neurosci. **11**, 423 (1988)
59. A. Muthuramalingam, S. Himavathi, E. Srinivasan, Int. J. Inf. Technol. **4**, 95 (2008)
60. B. Noory, V. Groza, IEEE CCECE 2003, p. 1861 (2003)
61. D.B. Strukov et al., Nature **453**, 80 (2008)
62. J.J. Yang et al., Nat. Nanotechnology **3**, 429 (2008)
63. Q. Xia et al., Nano Lett. **9**, 3640 (2009)
64. S.H. Jo et al., Nano Lett. **10**, 1297 (2010)
65. X. Wang et al., IEEE Electron Device Lett. **30**, 294 (2009)
66. Y.V. Pershin, M. Di Ventra, Phys. Rev. B **78**, 113309 (2008)
67. B. Linares-Barranco, T. Serrano-Gotarredona, Memristance can explain spike-time-dependent-plasticity in neural synapses, in *Nature Proceedings* (2009)
68. G.S. Snider, NANOARCH (2008) pp. 85–92
69. J. Borghetti et al., Nature **464**, 873 (2010)
70. A.V. Kolobov et al., Nat. Mater. **3**, 703 (2004)
71. K. Shportko et al., Nat. Mater. **7**, 653 (2008)
72. M. Wuttig, N. Yamada, Nat. Mater. **6**, 824 (2007)
73. S. Raoux, W. Welnic, D. Ielmini, Chem. Rev. **110**, 240 (2010)
74. V.G. Karpov et al., Appl. Phys. Lett. **90**, 123504 (2007)
75. S.R. Ovshinsky, Jpn. J. Appl. Phys. **43**, 4695 (2004)
76. S.R. Ovshinsky, B. Pashmakov, Mater. Res. Soc. Symp. Proc. **803**, 49 (2004)
77. S.R. Ovshinsky, in *E\PCOS* (2010)
78. T.C. Chong et al., Appl. Phys. Lett. **88**, 122114 (2006)
79. T.C. Chong et al., Phys. Rev. Lett. **100**, 136101 (2008)
80. T. Hasegawa et al., Adv. Mater. **22**, 1831 (2010)
81. Q. Lai et al., Adv. Mater. **22**, 2448 (2010)
82. C. Eliasmith, M.B. Westover, C.H. Anderson, Neurocomputing **44**, 1071 (2002)
83. P.D. Kuo, C. Eliasmith, Biol. Cybern. **93**, 178 (2005)
84. G.L. Gerstein, P. Bedenbaugh, M.H. Aertsen, IEEE Trans. Biomed. Eng. **36**, 4 (1989)
85. K.D. Harris, Nat. Rev., Neurosci. **6**, 399 (2005)
86. N. Kriegeskorte et al., Neuron **60**, 1126 (2008)
87. V.M. Sloutsky, Similarity, induction, naming and categorization: a bottom-up approach, in *A Neo-Constructivist Approach to Early Development*, ed. by S.P. Johnson (University Press Oxford, London, 2009)
88. V.M. Sloutsky, H. Kloos, A.V. Fisher, Psychol. Sci. **18**, 179 (2007)
89. T. Binzegger, R.J. Douglas, K.A.C. Martin, J. Neurosci. **24**, 8441 (2004)
90. G.L. Shaw, E. Harth, A.B. Scheibel, Exp. Neurol. Exp. Neurol. **77**, 324 (1982)
91. C. Koch, *The Quest for Consciousness: A Neurobiological Approach* (Roberts & Company Publishers, 2004)
92. D. Balduzzi, G. Tononi, PLoS Comput. Biol. **5**, 1 (2009)
93. G. Tononi, Biol. Bull. **215**(3), 216–216 (2008)
94. E. Basham, Z. Yang, W. Liu, IEEE Trans. Biomed. Circuits Syst. **3**, 321 (2009)
95. K. Chen et al., IEEE J. Solid-State Circuits **45**, 1946 (2010)
96. Z. Yang, Q. Zhao, W. Liu, J. Neural Eng. **6**, 046006 (2009)
97. Z. Yang, Q. Zhao, W. Liu, Neurocomputing **73**, 412 (2009)
98. V.B. Mountcastle, J. Neurophysiol. **20**, 408 (1957)
99. D.H. Hubel, T.N. Wiesel, J. Neurophysiol. **28**, 229 (1965)
100. D. George, J. Hawkins, PLoS Comput. Biol. **5**, e1000532 (2009)
101. K. Fukushima, Biol. Cybern. **36**, 93 (1980)
102. M. Reisenhuber, T. Poggio, Nat. Neurosci. **2**, 1019 (1999)
103. K. Ramanathan, L. Shi, T.C. Chong, COGSCI, 2010, pp. 1106–1111
104. S. Smale et al., Found. Comput. Math. **10**(1), 67–91 (2010)