NANTERO

White Paper

NRAM Carbon Nanotube Non-Volatile Memory

NRAM Carbon Nanotube Non-Volatile Memory… Can DRAM be replaced?

By Bill Gervasi & Rick Ridgley

Abstract: Nantero NRAM® is a non-volatile memory device deploying carbon nanotubes as the data retention element. These are fabricated on standard wafers using typical spin coat equipment and lithographic etching. The resulting cells operate over a very wide temperature range and are impervious to external effects such as radiation, shock, vibration, magnetism, etc. Memory arrays constructed using NRAM cells operate at read and write speeds comparable to mainstream DDR5 SDRAM. Cells may be constructed in 3D, achieving per-device densities exceeding SDRAM limits. Non-volatile NRAM improves on DRAM in the same applications by better performance from the elimination of refresh, and by lower power due to non-destructive core activation which eliminates the need for precharging. NRAM potentially becomes the universal memory solution, replacing both DRAM and Flash memory in many application.

About the Authors: Bill Gervasi is Principal Systems Architect for Nantero, and a chairman of the JEDEC international standards organization coordinating the development of memory and storage solutions for the computer industry. Rick Ridgley is Chief Scientist for Nantero and consultant to the Department of Defense on emerging and game changing technologies.

I. INTRODUCTION

Nantero was founded in 2001 to develop carbon nanotube (CNT) technologies for applications including use as a memory data cell. The technology was originally developed for use in aerospace and military applications in 2008, and the first devices were tested on board the Space Shuttle in 2009.

Nantero continued development of CNT for commercial applications targeting the mainstream data center and hyperscaler markets. The NRAM design is based upon the standard DDR5 SDRAM standard, and as a chairing member of JEDEC, Nantero has submitted specifications for a DDR5 NVRAM annex to standardize and document the unique features of these devices.

Nantero has built a number of test devices allowing characterization of the device performance including environmental conditions as well as basic cell reliability. In 2022, the testing achieved a requisite performance metric of 5-sigma separation at one million cycles for SET and RESET states of the devices, an important milestone for taking the device to mass production.

II. OVERVIEW OF USING CARBON NANOTUBES AS MEMORY

Carbon nanotubes, also known as CNTs, are among the toughest atomic constructions imaginable, comparable to diamond. They remain neutral to external effects such as heat, cold, magnetism, and radiation. Beneficially, CNTs exhibit a known resistance which is exploited to create a memory cell.

A CNT memory cell is constructed with a stochastic array of many carbon nanotubes1. Forcing CNTs to connect causes a change in cell resistance, and disconnecting causes the opposite change in cell resistance. This change is detected to create 1s and 0s of a memory storage element.

CNT cells are arranged into arrays which may be presented to the host memory controller with a custom interface, or as any of the standard memory interfaces including SDRAM.

CNTs are inherently non-volatile due to molecular binding forces that keep the nanotubes connected or separated. Data retention is measured in centuries, even millennia. The availability of high speed persistent memory enables a number of system level design choices from main memory to long term data storage that improve the performance of all computing systems.

Industry roadmaps indicate that SDRAM may hit a plateau in in per-device density of 32Gb per chip during the DDR5 lifecycle (2020-2028). NRAM shows no such limitation and is projected to exceed DRAM capacity; is NRAM a logical replacement for DRAM?

III. CONSTRUCTION OF CNT MEMORIES

CNT memory cell construction is relatively simple. A layer of carbon nanotubes is constructed between two electrodes. A filtering of the CNT length and diameter is used based on the target process node for the cell to ensure that hundreds or thousands of switchable nanotubes are available within each cell, depending on node. Cells have been built and tested at multiple production process nodes from 180nm down to 28nm, and also in the lab at 15nm, with modeling showing a clear path to switching even below 2nm.

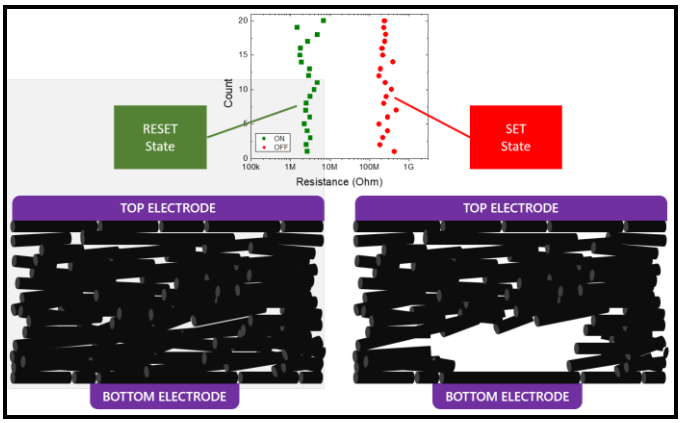

Figure 1: CNT Cells in RESET and SET States

As shown in Figure 1, the RESET state of each cell has a low resistance, and SET state a higher resistance. Device operating voltages are standard for SDRAM class devices: 1.8V and 1.1V. CNTs bond with adjacent nanotubes at the molecular level. Van der Waals (vdW) forces keep CNTs in either state without any external energy being needed. Energy is required to cross the vdW barrier in either direction. Data retention of CNT memory cells even under extreme conditions is rated in the hundreds of years; testing has shown operation from -55°C to +300°C.

Data retention of the CNT cells at +300°C is projected to be in excess of 300 years based on test data collected, applying standard reliability formulae. NRAM has been tested in space. It was aboard the space shuttle Atlantis during the Hubble repair mission where it performed flawlessly despite environmental stresses including shock, vibration, temperature extremes, plus high alpha and gamma radiation..

IV. MEMORY CELL APPLICATION FOR CNT

A. Optimizing the Memory Cell and Control Logic

Two common constructs of memory designs using CNTs are a transistor-per-resistor ("1T-1R") arrangement or a crosspoint arrangement. Figure 2 shows the distinction between these two arrangements.

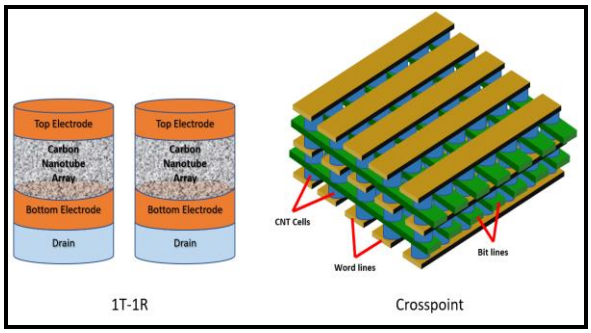

Figure 2: 1T-1R and Crosspoint Construction of CNT

With 1T-1R, the CNT cell is deposited directly onto a terminal of a switching transistor, much like a DRAM cell, and switched into a sensing amplifier circuit to detect the cell resistance. With a crosspoint, word lines and bit lines are directly connected to arrays of CNTs and act as the path for SET, RESET, and READ. These options are shown in Figure 2. The I-V curve of a CNT cell is advantageously non-linear such that it enables sensing the state of the selected cell without need for isolation diodes in the array. With a relatively low READ voltage, no evidence of read disturb has been seen in testing.

The manufacturing processes for 1T-1R and crosspoint are identical. The only difference is the connection scheme used between drivers, receivers, and the CNT elements. This allows for high speed storage elements such as registers, SRAM replacement, weight matrixes, or non-volatile FET controls to be deployed at the same time as slower bulk storage for data without adding manufacturing cost.

B. Performance Guidelines

Read and write performance are also critical parameters. Since switching is on the order of angstroms of distance, cells using CNTs switch between SET and RESET states in 5ns or less as shown in Figure 3. This is comparable to DRAM switching speeds, and makes CNT memory directly competitive with DRAM. Using performance as a guideline while specifying the word line and bit line lengths in the memory array, a 5ns core speed translates to DDR-class performance at the balls of the memory device: 45ns cycle time, 20ns access time, etc.

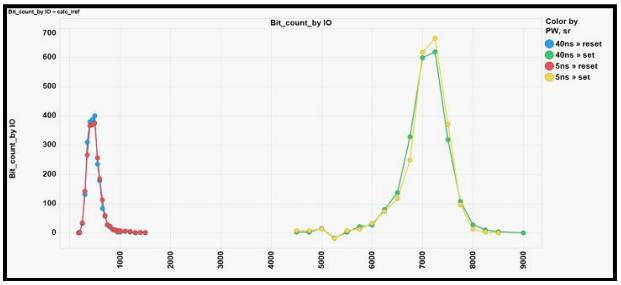

Figure 3: Core Performance Measurements to 5ns

C. CNT is Inherently Rad Hard

NRAM arrays have been tested on silicon substrates to 12MRad and $1E^{14}$n/cm with no statistically significant change in switching voltage or fabric resistance. No data was lost under these extreme conditions.

Each of six biased total ionization doses (TID) was supplied using cobalt 60 gamma radiation source targeting exposure doses of at least 2MRad(Si) each. Unbiased neutron displacement measurements of CNT logic was tested to $1x10^{14}$n/cm^2 and biased single event effect (SEE) testing with the linear energy transfer (LET) ion species ranging from 1.5MeV/mg/cm^2 to 120 MeV/mg/cm^2, the maximum required range for strategic applications.

The conclusion from this testing is that the environmental reliability of CNT memory cells is dependent on the underlying logic design and process, not the CNT cells.

D. The I/O Interface is Flexible

NRAM is a memory architecture deploying CNT cells for the storage array. NRAM presents this core to the system using standard I/O interfaces. DDR5 SDRAM is one such interface standard, as are LPDDR, GDDR, HBM, and others. These interfaces translate from the CNT core cell array to external addressing constructs such as bank groups, banks, rows, and columns. CNTs may be abstracted in other ways as well, including custom I/O interfaces, but also non-RAM-like interfaces.

E. CNTs Constructed in 3D

Fabrication of NRAM cells is typically done as a back end of line add-on process. The CNT slurry is applied using standard spin coat processes and equipment, then metal layers are sputtered over the CNT. Standard lithographic steps singulate the CNT cells into cylindrical pucks. These steps may be repeated as many times as necessary to achieve the desired memory density.
For example, in the construction of the DRAM-compatible design shown in Figure 5, carbon nanotube memory cells are constructed with 30nm diameter pucks on a 12-14nm logic process.

This yields a density of 8Gb on a 100mm² die. Additional layers are deposited on top of each other, with alternating metal layers for word lines and bit lines. Using this method, NRAM achieves the maximum capacity of the DDR5 interface, 32Gb per die, with four layers of CNT. Die stacking is also supported, yielding 512Gb in a 16-high 3DS stacked single package.

As a back end of line process, NRAM is also compatible with essentially any other manufacturing process including memory processes but also non-silicon and radiation-hardened processes. In each case, the maximum memory density achievable is a function of the size of the driver and receiver transistors and the design rules for metal pitch. The NRAM process combined with fine lithography memory processes promises to provide device densities in high gigabits to terabits per die in the coming decade.

In the DDR5 NVRAM specification submitted to JEDEC, a "row extension" feature is detailed that enables die densities from 64Gb to 1Tb per die without breaking the DDR5 protocol. This extension allows higher capacity NRAMs to be developed, such as on 7nm processes, without abandoning DDR5 compatibility.

F. NRAM is Lower Power Than DRAM

The non-volatile nature of the NRAM core provides significant power savings compared to SDRAM. For example, SDRAM cores have a "destructive activation", which means that when a row is read from the core to sense amplifiers, the contents of that row are destroyed. SDRAMs therefore must follow all activations with a "precharge" operation that restores the contents from the sense amplifiers to the core after reads or writes to the sense amplifiers are completed. In a typical DDR5 SDRAM, activation consumes 11% and precharge consume 21% of SDRAM power.

NRAM follows the SDRAM protocol – activate, read or write, precharge – but activations do not destroy the core content. Reads are performed directly to a small set of sense amplifiers, and writes directly modify the core. Precharge is treated as a NOP (no operation).

SDRAM cores, comprised of billions of capacitors, leak over short time intervals measured in milliseconds. As a result, SDRAMs require that every cell be read and rewritten to restore the desired voltage levels in an operation called refresh. This requirement holds whether the SDRAM is in use or in low power standby where the device is required to self refresh, burning considerable power.

NRAM requires no refresh (these commands are also considered NOP), and when the device enters standby mode, the NRAM only needs to maintain the device I/Os active to sense when standby mode is exited, conserving the vast majority of standby power. In fact, NRAMs can be

turned off completely for true zero power standby without losing data, though the interface with the host system may require recalibration upon exit from zero power standby.

SDRAM rows are 8Kb in size due to their internal 2D array structures. Therefore, every activation and precharge operation requires reading and rewriting all 8Kb, even if only 64 bits of the row are accessed. This means that SDRAM cores are only 0.8% efficient in terms of power for work performed. NRAM exploits activate-in-place, read-in-place, and write-in-place capabilities of its non-volatile 3D core to reduce this penalty so that only 64 bits of the core are operated on at a time.

Combining these various power savings features, NRAM can reduce memory power over SDRAM by between 21% to 50%. The majority of power burn for NRAMs comes primarily from the requirements of the DDR5 I/Os, not the core power.

G. NRAM Thermal Characteristics are Beneficial

Carbon nanotubes are essentially rolled up carbon, the same material as diamonds, so it is not surprising that the thermal characteristics are very similar to diamond. As shown in Figure 4, carbon nanotubes at 800W/mK are second only to diamond's 1000W/mK. Adding CNT to a design can act as an excellent heat spreader.

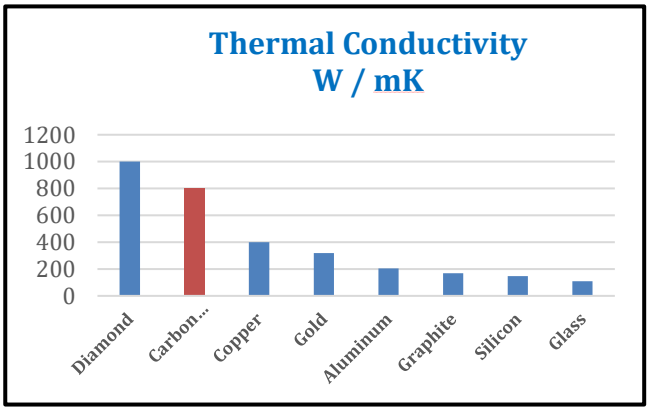

Figure 4: Thermal Characteristics of Various Materials

H. NRAM is Inexpensive

No new memory technology can compete without also providing a cost effective solution. NRAM is an inherently low cost technology. CNT is constructed using low-cost and readily available carbon sources. NRAM slurries are applied to wafers using standard spin coating equipment and baking ovens, already installed in all fabs, for the deposition of the CNT layers,

and standard sputtering for additional metal layers. Also standard are the lithographic manufacturing steps for patterning and etching. No new equipment is required to install CNT processing in a fab.

V. DDR5 IMPLEMENTATION

A. NRAM Used as DRAM Replacement

Nantero's DDR5 NRAM design, shown in Figure 5, employs the crosspoint CNT arrangement discussed in the previous section to create a nonvolatile DDR5 SDRAM drop-in compatible device. From the outside looking in, a DDR5 NRAM provides the same signals, timing, and electrical characteristics as a JEDEC standard DDR5 SDRAM. This includes support for the CID inputs that allow for die stacking using through-silicon vias to provide 16-high chips stacks. With a per-die capacity of 32Gb, this allows 512Gb or 64GB of non-volatile memory per package. To increase yield and reliability, DDR5 NRAMs incorporate on-the-fly error correction. Due to the fast switching speed of the CNT cells, ECC can be handled on-the-fly without violating DDR5 timing.

Redundancy and replacement of bad bits also increase device yields. The Built-In Self Test and Mapping CAM blocks shown in Figure 5 allow for detection of bad bits in the NRAM array and replacement using spare blocks.

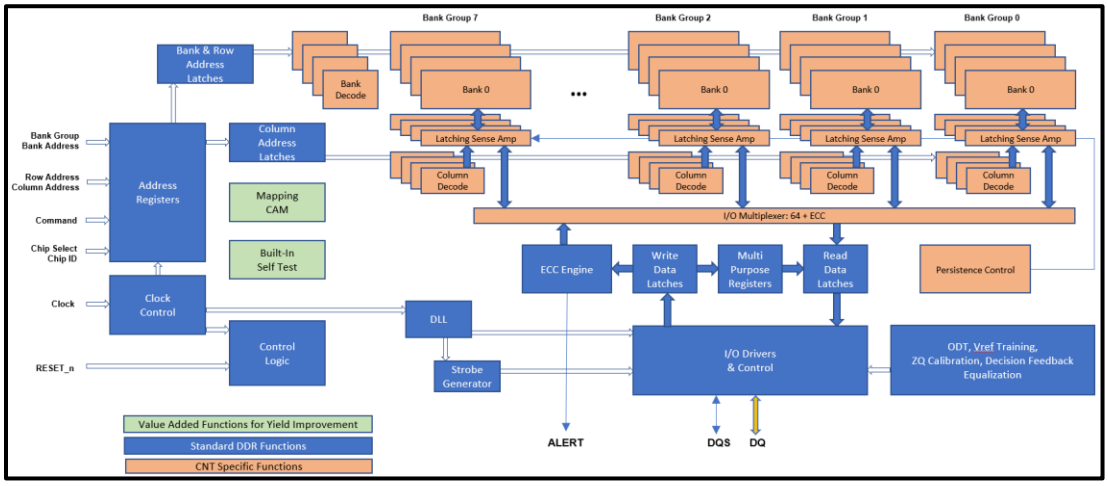

Figure 5: DDR5 NRAM Block Diagram

B. NRAM is Immune to Rowhammer Attacks

Rowhammer is an attack mechanism used by malicious software to break into or crash computing systems. Rowhammer takes advantage of two characteristics of SDRAM design: 1)

the capacitive cells that hold data lose their charge over short periods of time measured in milliseconds, hence why SDRAMs need refresh, and 2) the 2D nature of SDRAM core layout makes these cells susceptible to crosstalk. Rowhammer attacks perform thousands of activations on rows in the SDRAM with the intent to cause crosstalk that will change the contents of nearby rows of memory, thereby handing control over to planted virus code. SDRAMs are particularly susceptible to these attacks in the second half of a refresh period where the cell voltage has drooped significantly. Data centers often are forced to double or quadruple refresh rates to reduce this sensitivity, with severe impacts to system performance and no guarantee that the protection is sufficient to prevent break-in.

By design, NRAM solves the rowhammer sensitivity problem in two ways. First, the persistent nature of the NRAM cell means that the stored data values do not droop over time. Second, the internal structure of an NRAM crosspoint array is protected from crosstalk through judicious assertion of inhibit voltages that prevent sneak currents and crosstalk from modifying adjacent rows of memory.

VI. BEYOND VON NEUMANN

Non-von Neumann architectures are increasingly getting attention as various forms of artificial intelligence and machine learning architectures emerge as mainstream solutions in applications ranging from ordering pizza to analyzing financial trends to mining security data. While there are far too many to list exhaustively (and avoiding the semantic wars over the term "artificial intelligence"), Figure 6 shows graphically a few variations of AI cores used to accelerate applications from hyperdimensional vector weighting units to analog multiply-accumulate units and beyond.

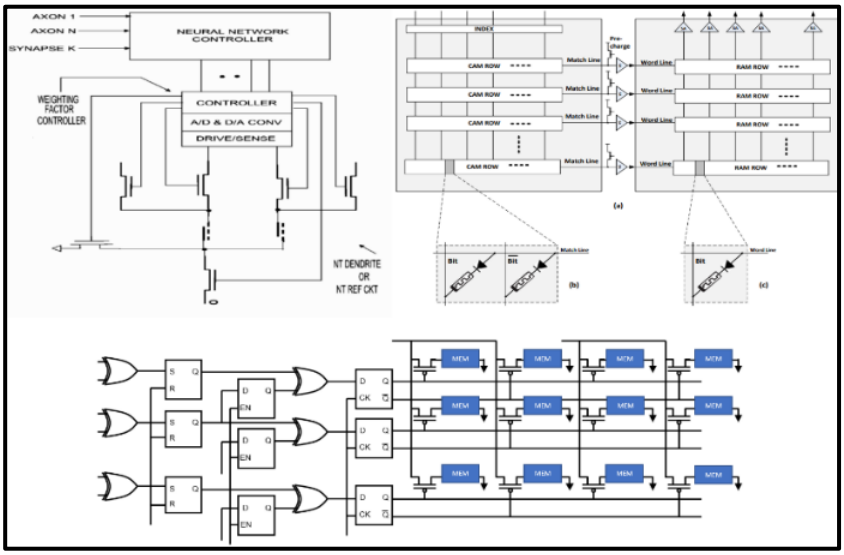

Figure 6: Some Varieties of AI Cores

One common aspect of most AI solutions is buried data. The storage element is best contained within the processing element. However, increasing use of back propagation in AI algorithms means that learned and corrected data is stored in the computation circuits themselves. This discourages the use of volatile memory solutions like DRAM or SRAM because of the vulnerability of data loss on power failure. AI architectures typically give up significant device and system bandwidth to checkpoint critical data to external permanent storage such as SSDs. Incorporation of non-volatile high performance NRAM cells into an AI architecture not only solves the problem of potential data loss, it frees design engineers from constraints imposed by volatile AI storage elements. Written data can be kept in place as long as needed, and mechanisms for checkpointing, if necessary at all, can be tuned to deal with catastrophic system failures as opposed to more frequent power failure concerns.

Nonvolatile memory in an AI engine also assists greatly with power consumption and heat issues. A shared memory, such as an HBM device, requires significant power to access to transfer data into the processing core, but shared HBM is often used because it simplifies the data checkpointing requirements. Conversely, a non-volatile memory cell in the processing element eliminates this high power data fetching operation. For maximum power savings, AI engines with integrated non-volatile memory can turn power off completely to unused processing cores, or between operations, without risk of losing data.

AI engines tend to be capacity limited as well. The more data they can access locally, the better they can operate. The density trend of NRAM into terabits per device aligns exceptionally well with the increasing capacity demands of the AI market.

V. Conclusions

Can NRAM replace SDRAM? All indications are that it is poised to do so. NRAM offers the same performance as SDRAM, while providing significant benefits including data persistence, higher performance, lower power, and a roadmap to device capacities exceeding the SDRAM roadmaps. For over 50 years, system architectures have been based on the assumption that main memory is volatile. For the first time in decades, that may change.

--bg, rr