Coughlin Associates

Machine Learning Enables Longer Life High Capacity SSDs

NVMdurance White Paper

Tom Coughlin
Coughlin Associates

www.tomcoughlin.com

Executive Summary

Machine learning can be used in flash memory devices to optimize storage capacity, endurance and data retention. The way that data is stored in flash memory devices leads to wear and limitations in data retention with repeated erase and program cycles and even with multiple reading of cells. These issues become even more important as the number of bits per flash memory cell increases. The tuning of registers in 2D as well as 3D flash memory can be used to optimize flash wear, storage capacity and data retention; but as the number of these registers increases, this task becomes impossible to do manually. The NVMdurance Pathfinder product uses machine learning to automate the optimization of the flash memory register trade-offs for real flash memory products. The company's Navigator product, running on an SSD, implements these optimized register settings with drive use. This approach creates longer lasting flash memory products and will become pervasive in the industry. It provides a competitive advantage for those who implement it, and a disadvantage for those companies who don't.

Introduction to Flash Memory Storage Capacity, Endurance and Data Retention

We will discuss the trade-offs that are made between endurance, retention and storage capacity in the design of controllers for flash memory and for various applications. But first let's discuss how flash memory is written and erased and how this leads to endurance and data retention issues, which typically get worse as the storage capacity increases.

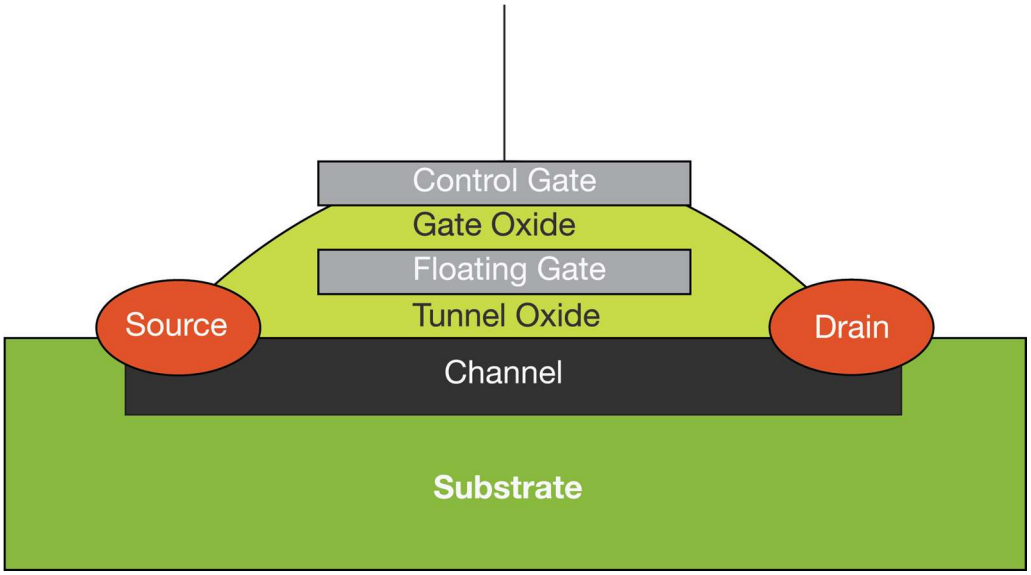
Flash Memory stores information as electrons on a floating gate or charge trap in a flash memory cell transistor. In a floating gate these electrons are quantum mechanically tunneled (Fowler-Nordheim tunneling) through an insulating layer to a conductive gate that is isolated within the insulator (see **Figure 1)**. When a bit is programmed, electrons are stored upon the floating gate. This has the effect of offsetting the charge on the control gate of the transistor.

Fowler-Nordheim tunneling requires a high voltage (usually between 7-12 Volts) to be placed between the source and the control gate of the transistor. If the voltage is sufficient, the electrons "tunnel" through the gate oxide layer and come to rest upon the floating gate.

In NAND and NOR flash technology, a single very large transistor erases all the transistors in a subarray called a "block". This provides a significant cost savings to flash chip designers who don't need to be able to individually erase each bit or byte of the memory.

Tunneling electrons migrating through the tunnel oxide sometimes causes difficulties. It is inevitable that electrons will get trapped in the tunnel oxide with erasing and writing. These electrons, once trapped, cannot be removed but they can free themselves as a function of time and temperature. This will impact the operation of the cell to a certain degree, depending upon the number of electrons trapped in the oxide: if a lot of electrons are trapped, then there will be a big impact, but a low number of electrons are not likely to cause much impact at all. As the number of trapped electrons in the oxide gets too large it increases the apparent voltage on the floating gate and may give rise to bit errors or erase failure.

Figure 1. Floating Gate Flash Memory Cell



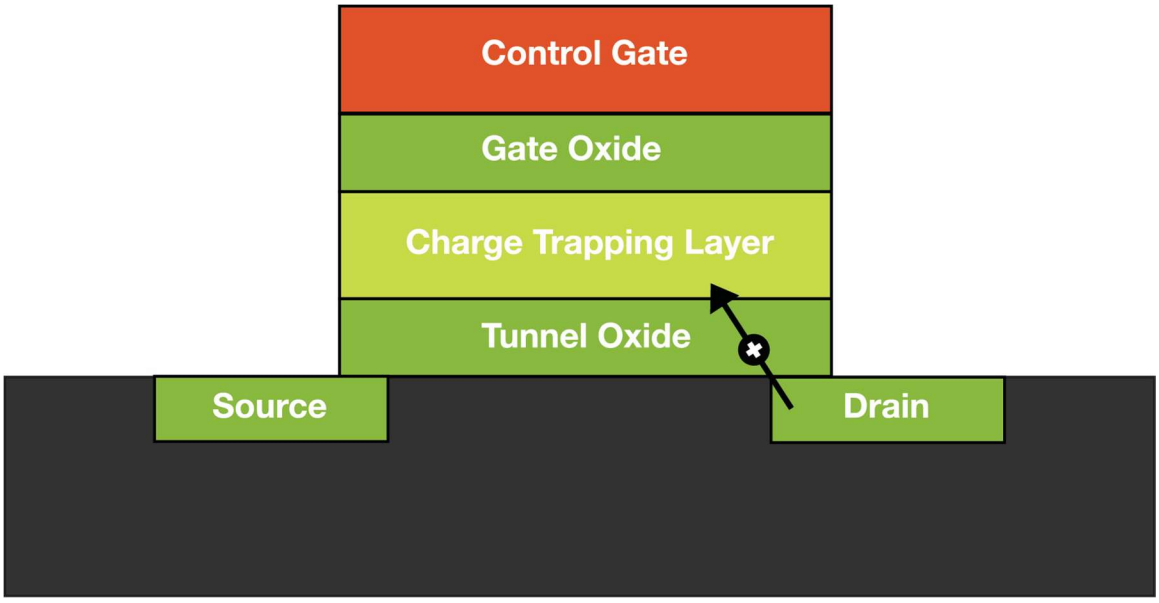
The tunneling of electrons also weakens the insulator, making it harder to keep electrons on the floating gate. As the electrons leak off the gate over time this changes the voltage on the floating gate and also leads to bit errors. The amount of leakage directly affects the length of time that the correct data can be stored on the floating gate (the retention time). The more times the cell is programmed and erased, the weaker the insulator layer becomes and the lower the retention period. The damage to the isolation insulator in the flash memory cell is referred as cell wear.

The number of electrons trapped in any one cell's tunnel oxide is a function of how the chip is made and more importantly on how many times that particular memory cell has been erased and rewritten. A specification has thus been devised to recommend a maximum number of erase/write cycles a memory cell on a chip can withstand before a failure is likely to occur, and this specification is called the chip's "Endurance".

In a charge trap design, electrons are stored in an insulating layer rather than on a conductor, as in a floating gate (see **Figure 2)**. Charge traps can use thinner insulating layers leading to faster erase speeds.

The electrons stored on a flash memory cell gradually tunnel back out of the floating gate or charge trap. As the charge stored in the cell declines, eventually the signal representing the data stored in the cell declines, resulting in a reduced signal to noise ratio (SNR). When the SNR gets too low the data can no longer be recovered from the cell without using very sophisticated and expensive error correction codes (ECC) and data recovery methods.

Figure 2. Charge Trap Gate Flash Memory Cell



Data retention is the length of time that flash memory cells can store a recorded bit before that data becomes irrecoverable. Data retention is highest when a flash memory product is new and declines over the specified life of the flash memory device. Data retention is closely coupled to flash cell endurance. Data retention specifications for SSDs are typically given for a flash memory cell at the end of its specified endurance life[1]. For enterprise SSDs, data retention at end of product life should be at least three months, and for client SSDs at the end of life, the data retention should be at least one year. Note that there may be other retention requirements for particular applications.

Flash Memory Types

There are two kinds of flash memory, NOR and NAND. The two terms are names of types of logic gates, the negated "or" function and the negated "and" function. The big difference between the two types of architectures is real estate. NAND has a significantly smaller die size than does NOR. This translates to significant cost savings.

These cost savings come with a trade-off. NAND does not behave like other memories. While NOR, SRAM and DRAM are random-access devices (the "RAM" part of DRAM and SRAM stands for "Random Access Memory") NAND is part random and part serial. Once an address is given to the device, there is a long pause, then that address and several adjacent addresses' data come out rapidly.

NAND cells are susceptible to bit errors. In order to correct these errors, NAND makers use error correction technology. For instance a 256-bit sector may have an additional 16 bits

[1] Addressing Data Retention in SSDs, Jon Tanguy, Micron, https://www.micron.com/about/blogs/2015/may/addressing-data-retention-in-ssds

added on for error correction. These check or parity bits are programmed with a highly-compressed code that indicates what the data in those other 256 bits should look like.

Every time that a sector is programmed, corresponding parity bits must be calculated and stored along with the original bits. A controller is usually given this task. This is one of the reasons that NAND most often ships coupled with a controller chip, although there are cases where the functions of the controller are embedded within another processor within the system.

During a read, the data bits in a sector cannot always be trusted, so they are never simply read from the NAND into the system. Instead, the 256 bits in the sector are combined with the 16 parity bits in an error detection and correction engine, which presents a sector's worth of corrected data to the system.

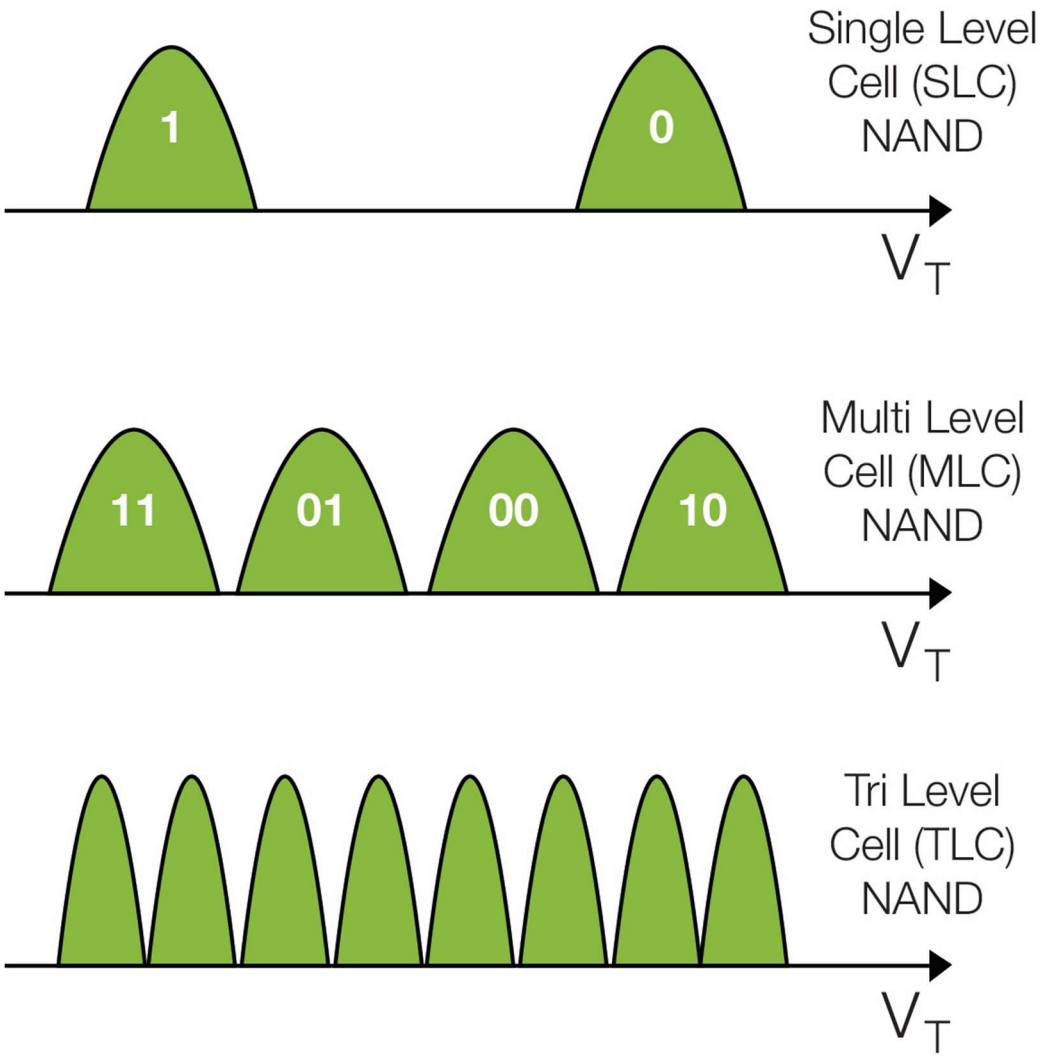
Flash memory wear, as discussed earlier, is managed by the flash memory controller. This is a microprocessor that controls the operation of the flash memory cells. The wear is managed by spreading the writing across all the flash memory cells as well as changing the characteristics of the program and erase cycles of the cells over the life of the device.

Along with wear leveling, the controller in a NAND-based system is responsible for bad block management. The controller, as a part of its data correction responsibilities, keeps track of how much error correction is performed on each sector of the NAND. If the number reaches some predetermined high level, that sector is marked as "bad" and will no longer be used by the controller. In NAND with bad block management, as blocks get taken out of the system, the overall capacity of the NAND device slowly diminishes.

Flash manufacturers increase the capacity of flash chips by making the transistors smaller using finer device lithographic features. Besides increasing the number of cells in the device, the number of bits per cell can be increased as well. A single bit per cell is called a Single Level Cell (SLC). If there are two bits per cell it is called a Multilevel Cell (MLC). If there are three bits per cell it is called a Three Level Cell (TLC). Four bits per cell is a quad level cell (QLC). We show how the same voltage span in a flash memory cell can be divided into multiple levels to make an MLC and TLC flash in **Figure 3**.

In order to store 2 bits per cell in MLC, the four different states represented by those two bits must each be present (00, 01, 10, & 11). You can do this by storing four voltage levels (or threshold voltages) on the floating gate. Instead of storing a high charge on the floating gate for a logical 1 and a low charge for a logical 0, using the halfway point as the decision level between a 1 and a 0, you can store four charge (voltage) levels: Nothing, one third, two thirds, and full, each representing one of the four states.

Figure 3. Voltage Levels for SLC, MLC and TLC Cells



When an MLC cell is read, the reading circuit has to discern between four small voltage levels rather than two big ones. Digital chips are noisy, and the more noise you have in a system, the more difficult it is to discern small voltages apart from one another. This challenge has caused leading flash suppliers a good share of headaches. One part of the solution is to give the chip time for the noise to settle down before making a decision. This slows the chip down, which is why MLC chips are usually slower on read cycles than their SLC counterparts.

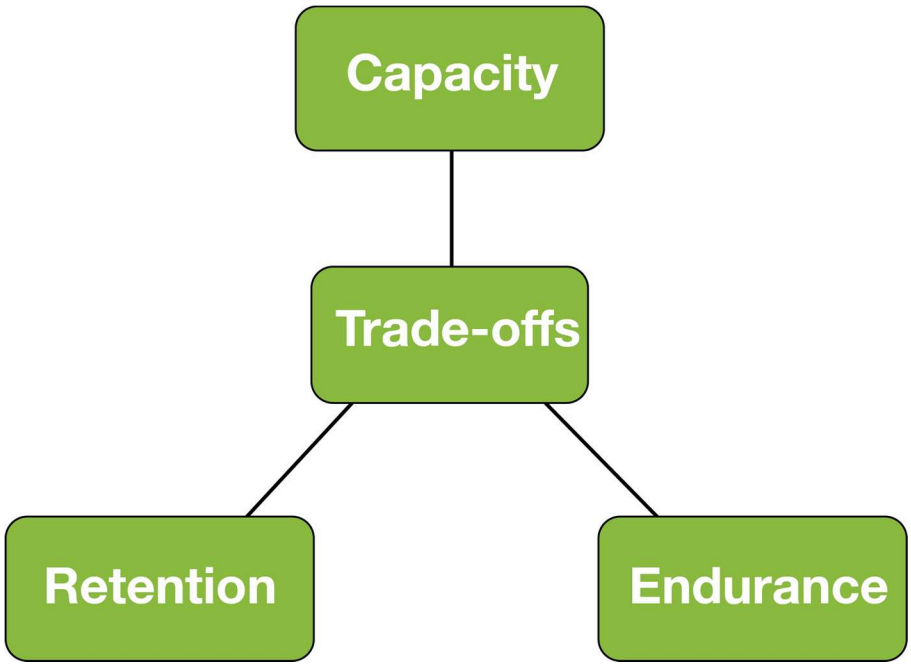
Somewhat similarly, writing into MLC chips is a bit more daunting than writing into SLC chips. Across the chip, different cells behave differently, and relatively sophisticated state machines (tiny little computers) are used first to put a small amount of charge onto a floating gate, in order to see if the floating gate's charge is near the optimum charge for the cell to represent one of the four levels. It then continues iterating on this process until it is done. Usually a smaller current is used to give better control over the charge, and this slows down programming.

As we increase from 2 bits per cell to 3 bits per cell and even 4 bits per cell we divide the number of charge levels on the floating gate into more levels. The SNR of each level declines as the number of levels increases. Besides slowing the write time of the cells, the lower SNR makes the memory more sensitive to adjacent cell disturbs where part of the charge on one flash cell leaks into an adjoining cell. As the levels of charge on the floating gates becomes smaller and smaller, the impact of a few electrons migrating from one floating gate to another becomes more and more significant.

Tuning of Flash Memory Parameters

As has been discussed above, there are ways to trade off the capacity of a flash memory device, versus its endurance and its data retention (see **Figure 4)**.

Figure 4. Trade-offs between Storage Capacity, Data Retention and Endurance



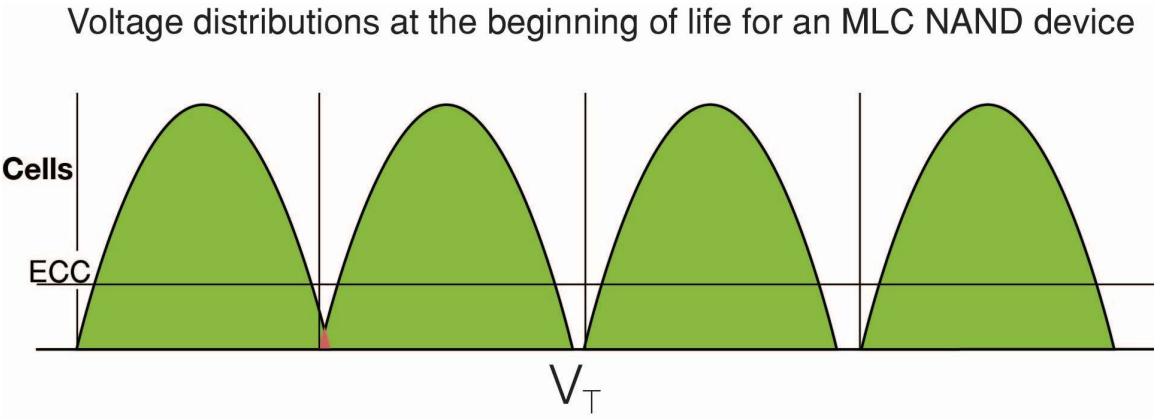
This can be done partially in the initial design and manufacturing of the device, but these characteristics can also be optimized for the application by changing the settings of various registers in the controller that control the operation of the flash memory over time. Let's look at some of things these registers control. The number of registers can be as few as 50 for 2D (planar) flash and over 1000 for 3D flash memory.

For instance, the insulator that isolates the stored or trapped electrons in a flash memory cell requires a voltage to allow electrons to be tunneled onto the gate or trap when writing or removed from the gate or trap when erasing. When the flash memory cell is new, the voltage required to move these electrons (particularly for erase) is less than the voltage required later in the life of the memory cell. A higher voltage applied early in the life of the cell will cause the insulator to start to break down and thus the cell will wear out faster.

The registers that control the voltages applied during writing or erasing the cells can be changed over the life of the cell to increase the endurance of the cell.

There are also registers that control the voltage thresholds that determine the bits recorded on a multi-level cell (see **Figure 5)**.

Figure 5. Example of Two Bit Per Cell Voltage Thresholds



The optimal voltage thresholds change over the life of the flash memory, again related to the wear of the cells over time. Also as the lithographic features decrease these voltage regions start to run into each other. As a consequence of this movement and with finer cell lithography the Bit Error Rate (BER) suffers and the Error Correcting Codes (ECC) running on the controller must be more robust, as well as applied more often, which negatively impacts overall device performance.

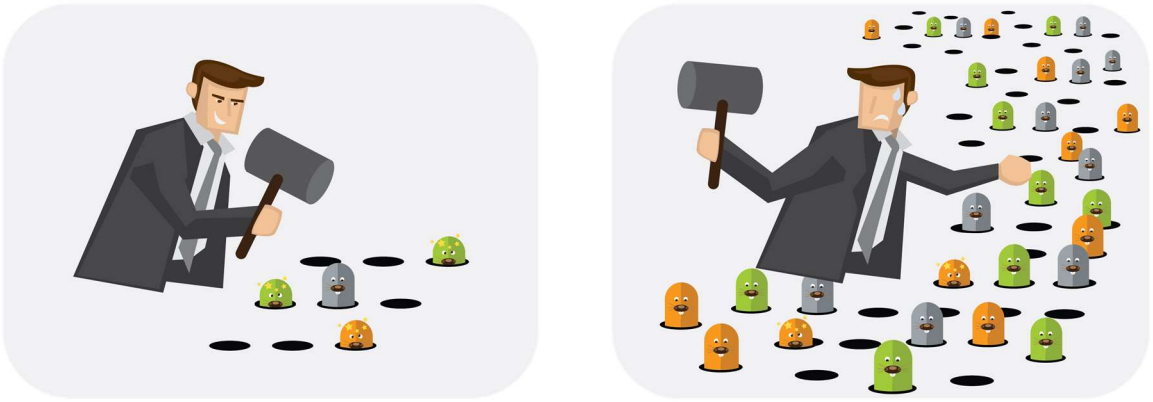
As these thresholds also get closer to each other and overlap, the number of bits per cell increases, so changes in these thresholds over time can impact the storage capacity and the ability to recover the stored bits. These changes may be caused by change in the individual cells over time, as well as issues such as noisy neighbors and write and read disturbs, that can cause errors. Note that the TLC threshold control over the life of the device is much more critical than MLC threshold control.

Successfully recovering the recorded bits in a cell is important to achieve the required storage capacity and also relates to the data retention of the cells. If the registers that control these threshold voltages change with time, data recovery is improved for a longer period of time, making the effective data retention of the memory cells longer.

There are other flash memory registers that control the amount of time that the voltages are applied, which is related to the number of electrons moved to the storage unit. There are also registers that control the ramp of the voltage application, as well as the number of retries that are required before a cell is declared dead and then de-allocated from working storage. Each voltage level in a multi-level cell has one or more registers associated with it.

Planar flash memory devices generally have 50-100 8-bit registers. With the additional layers of cells in 3D flash, the number of registers is much greater, by an order of magnitude or more. As the number of layers increases the number of registers will increase further. Manually optimizing the 100 registers in a planar flash memory device is daunting. Manually optimizing the 1000+ registers in a 3D flash device is impossible as suggested in **Figure 6**.

Figure 6. Manual Tuning of Too Many Manual Registers is Impossible



As a result, tuning or trimming the registers in modern flash memory devices, in order to optimize the applications requirements for storage capacity, endurance and data retention, will require automated tools using machine learning technology.

Using Machine Intelligence to Optimize Flash Memory

The ability to tune or trim flash memory registers more quickly results in faster yield improvements during product ramp and also results in fewer issues in the field. Automated tuning that adjusts these registers over the life of the device also provides for selecting combinations of parameter settings optimized for specific use cases. These combinations of parameter settings will increase the ability of a flash memory manufacturer to get product to a broader market faster and improve their revenue and profitability.

NVMdurance

NVMdurance, a young company from Limerick Ireland, has pioneered the use of machine learning to determine the optimal flash memory registers over the life of the flash memory device.

Machine learning can learn patterns and trends from data sets that can then act as a model for the operation of a real system. Such a machine-learning model can then be interrogated and searched to find the optimal ways to use the results of this model, depending upon constraints and goals. In this case the goal is to optimize the combination of storage capacity, endurance and data retention, depending upon the product use case.

Pathfinder

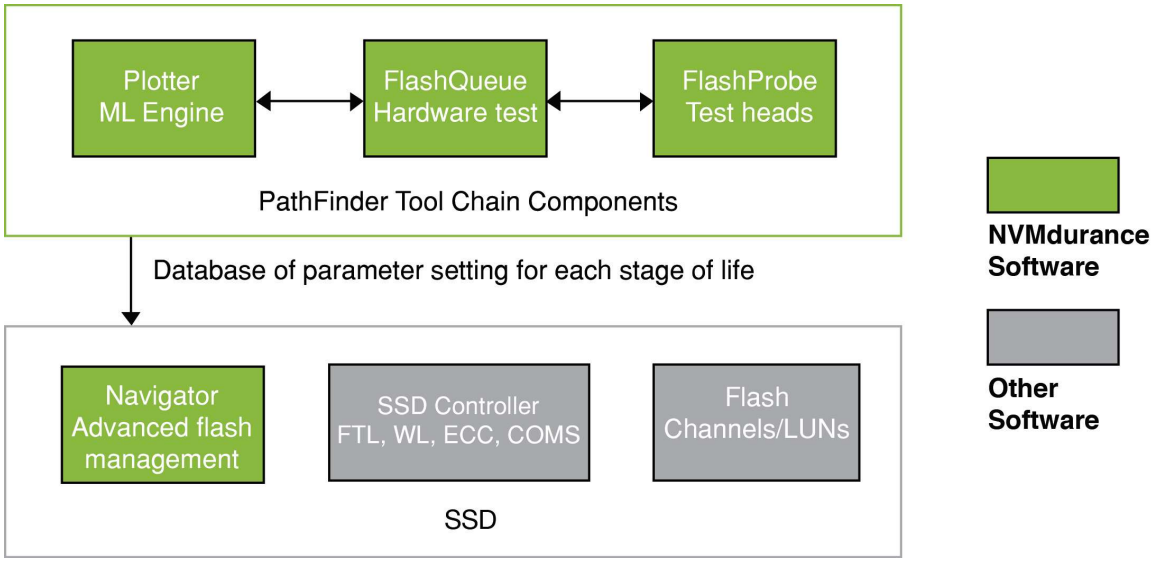
There are two important software-only elements in the NVMdurance solution. These are called Pathfinder and Navigator. Pathfinder is used to determine the optimal register settings during the life of the flash memory device in a laboratory setting. Navigator sits on the SSD in the field to control the application of these register settings depending upon use and health monitors.

The Pathfinder algorithm determines the search space and generates initial candidate solutions, a candidate solution being a set of potential flash register values. It then tests these initial candidates in hardware. It builds a predictive model using the hardware results and uses this model to predict better candidate solutions. These are then tested.

If the resulting candidate passes all tests, then it is added to the volume set that will eventually be placed in the controller registers for this device type. This process is repeated as many times as needed to find more useful candidate solutions.

Figure 7 shows a block diagram of the Pathfinder operations and the Navigator sitting on the SSD.

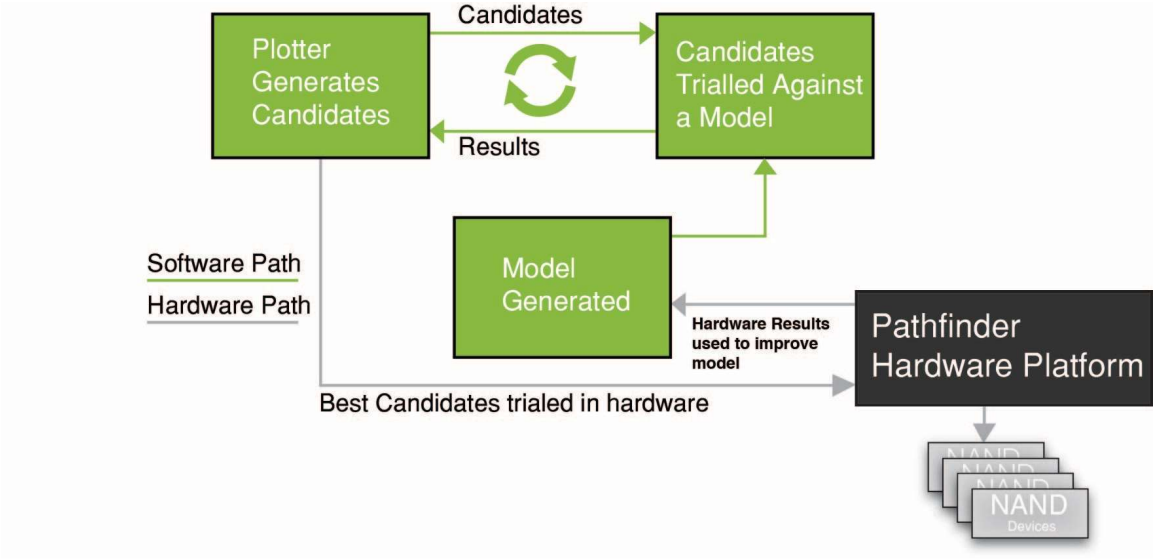
Figure 7. Pathfinder Elements and Navigator Running on the SSD



Plotter is the Machine Learning Engine that integrates test data generated on flash memory devices of a given type made by a given manufacturer in order to tune the register setting that will be used by the Navigator program running on the SSD over the life of the SSD. The FlashQueue program schedules hardware tests in the laboratory and performs online search and data collection to feed Plotter. FlashProbe is test head firmware that runs on flash memory devices to gather test data for FlashQueue.

Plotter works with a set of rules, such as shown in **Figure 8** that it uses to find the optimal register settings that will meet these rules.

Figure 8. Pathfinder Generation of Flash Memory Model



Plotter works with data from the FlashQueue and FlashProbe software running on the Pathfinder Hardware Platform, in order to iterate test trials run on millions of virtual flash devices. Plotter leverages the results of these trials to determine a model for the flash memory device over time. Note that there will be at least 3 iterations between Plotter and the measured flash memory data and as many as 10 iterations depending upon the goals for the device.

For instance, if you have a 1,000 erase/write cycle endurance flash memory chip, with a few iterations you can create a plan to accomplish 2,000 erase/write cycles. Many more iterations would be required to accomplish a 15,000 erase/write cycle endurance for the same flash memory chips.

FlashQueue consists of networked components used to schedule candidate testing using FlashProbe and collecting data such as BER and timing as shown in **Figure 9.**

Figure 10 shows the operation of FlashProbe. This is a single board computer containing standard flash read/write/erase operating grammar (e.g. per the ONFI spec), extended test mode grammar via an encrypted AFT (Abstract Flash Trimming) driver, temperature control for the test head and a timing measurement capability.

Figure 9. FlashQueue Operation

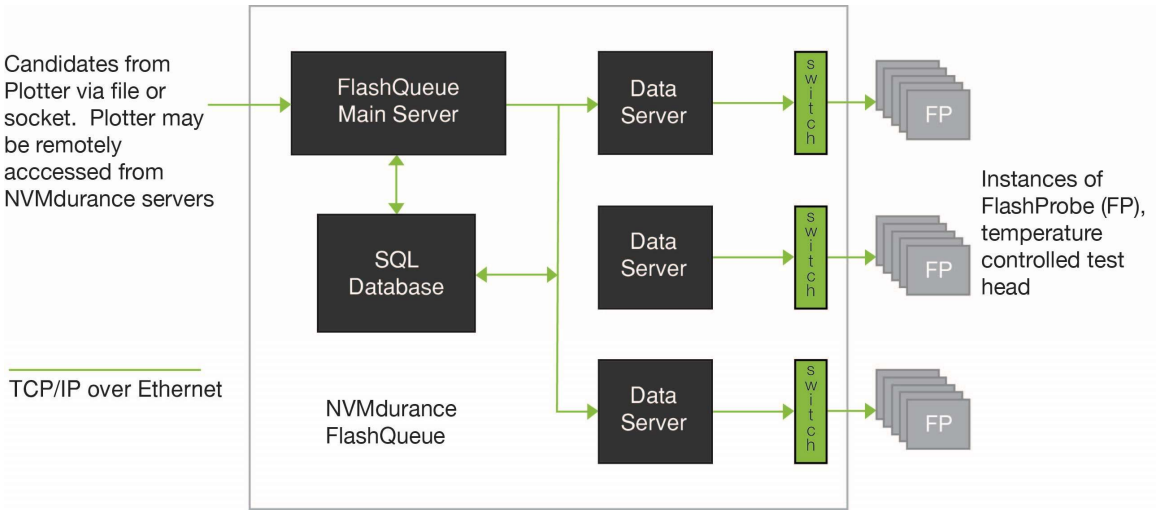
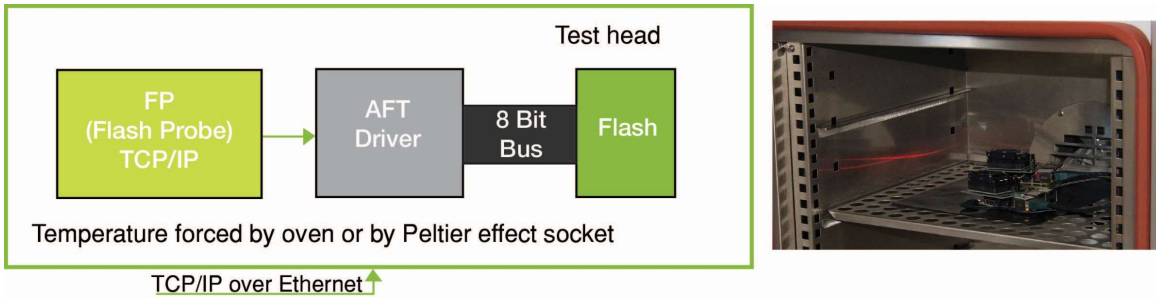


Figure 10. Operation of FlashProbe

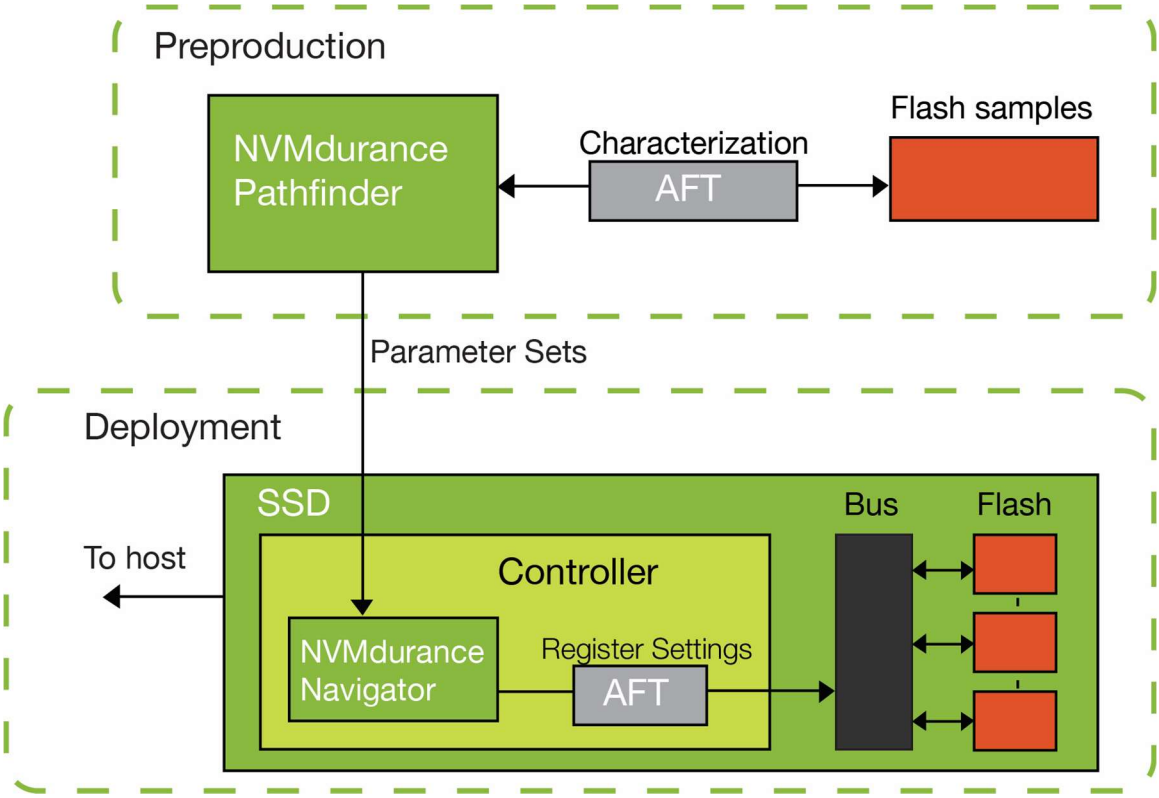


For NVMdurance (or an SSD controller firmware programmer) to change flash register values they need to have access to "test modes" that flash vendors are very loath to disclose. NVMdurance has mitigated this need by designing (and patenting) the Abstract Flash Trimming concept. This places an encrypted layer between the programmer and the flash device. It allows all of the functions without disclosing the methods.

Pathfinder can create an optimal model (set of flash register settings over time) that meet the machine learning rules and create optimal performance to meet expected usage for the flash memory device (including the full range of manufacturing variations of these flash memory devices). This optimization may make trade-offs between storage capacity, endurance and data retention.

The Abstract Flash Trimming file is created during the Pathfinder optimization and is encrypted. The AFT file is abstracted from the actual silicon operation. It records how registers interact with each other, not where they are or how they are accessed or stored on the SSD. This encrypted file is then unpacked and run on Navigator running in the product SSD as shown in **Figure 11**.

Figure 11. AFT Created in Pathfinder Runs on Navigator to Optimize SSD Capacity, Endurance and Data Retention.



Navigator periodically measures the "health" of the flash chips (a combination of BER, read, write and erase timing, plus historical profiling) in the field to determine when to change the SSD registers, corresponding with a new stage in the life of the SSD.

Navigator can also use the embedded features of the controller to further extend the life of the flash. For instance, if higher level of ECC is available, Navigator can extend the time that it can use weak write currents on the flash cells and thus reduce the rate of insulator damage growth. Also weak or compromised blocks can be rested or deleted from the overall population so they don't impact the setting on the healthier blocks.

Flash Memory Machine Learning Will Become Pervasive

Flash memory wears out with use, so the data stored in flash memory cells will disappear over time. Accomplishing a long useful life in terms of endurance, as well as the specified data retention time, along with the highest storage capacity, requires constant measurement and control of many parameters in the flash memory cells.

These parameters, or registers, must change over the life of the device to optimize the product life. Optimizing these register values over time is difficult to accomplish manually

for planar flash. With the increase in the number of cell registers in 3D flash, finding the best flash memory settings over time is impossible to do manually.

NVMdurance uses machine learning from iterative measurements on representative flash memory hardware with its Pathfinder software to determine the appropriate sequence of register settings that can be applied over time, based on degrading flash memory cell health. The company's Navigator software, running in an SSD, then autonomically uses these predetermined Pathfinder register settings to optimize the endurance and/or data retention for SSD in its actual usage / application environment.

Reducing the time to optimize flash memory register settings will result in faster production yield ramps for flash devices and likely fewer problems in the field. This optimization can also make it possible to create TLC or even QLC rather than MLC flash devices for many demanding applications, without resorting to expensive LDPC (Low Density Parity Check) error correction or more powerful controllers. LDPC error correction is more complicated than more common BCH (Bose, Chaudhuri and Hocquenghem) error correction code and requires more processing capability, which may reduce the overall chip performance. The NVMdurance technology can also avoid runtime impact and bad tail latency (long latencies in a distribution of operations) due to read retries. This optimization can furthermore compound any endurance gain achieved by some other means, such as more powerful ECC or overprovisioning.

These advantages make automated flash memory tuning, using machine learning, attractive and should result in this approach becoming pervasive in the industry, particularly as the industry moves to 3D flash.

About NVMdurance

NVMdurance was founded in Ireland in 2013 when the ADAPT research group and the test equipment maker Evolvability Ltd were brought together at the National Digital Research Centre in Ireland. The technology is the result of 15 years' foundry work on flash memory endurance optimization by the founding team.

The company's core IP is a set of NAND flash optimization techniques that have now been implemented in software. The team has been doing parameter discovery and memory characterization since 2000, initially with SLC NOR, later with SLC NAND, then MLC, then TLC and laterally, 3D NAND. This has all been done with silicon fabricators and major data storage manufacturers.

Originally applied manually in flash optimization work, NVMdurance has moved to a fully autonomous on-controller system, NVMdurance Navigator. Navigator constantly monitors the condition of the SSD and automatically adjusts the control parameters in real time. Prior to deployment of the flash memory, offline, a great deal of compute-intensive pre-processing is done by NVMdurance Pathfinder, a custom-built suite of machine learning techniques.

After extensive successful trials with multiple flash manufacturers NVMdurance is now announcing customer wins, expanding its team and focusing on the success of the early deployments of its technology.

About the Author

Tom Coughlin, President, Coughlin Associates is a widely respected digital storage analyst as well as business and technology consultant. He has over 35 years in the data storage industry with engineering and management positions at high profile companies.

Dr. Coughlin has many publications and six patents to his credit. Tom is also the author of Digital Storage in Consumer Electronics: The Essential Guide, which was published by Newnes Press. Coughlin Associates provides market and technology analysis as well as Data Storage Technical and Business Consulting services. Tom publishes the *Digital Storage Technology Newsletter, the Media and Entertainment Storage Report*, the *Emerging Non-Volatile Memory Report* and other industry reports. Tom is also a regular contributor on digital storage for Forbes.com and other blogs.

Tom is active with SMPTE, SNIA, the IEEE (he is Director for IEEE Region 6 and active in the Consumer Electronics Society) and other professional organizations. Tom is the founder and organizer of the Annual Storage Visions Conference (www.storagevisions.com), a partner to the International Consumer Electronics Show, as well as the Creative Storage Conference (www.creativestorage.org). He is also the general chairman of the annual Flash Memory Summit. He is a Senior member of the IEEE, Leader in the Gerson Lehrman Group Councils of Advisors and a member of the Consultants Network of Silicon Valley (CNSV). For more information on Tom Coughlin and his publications go to www.tomcoughlin.com.