ADVANTEST®

# Why Next-Generation NAND Flash Requires a Dedicated Test Solution

By Ken Hanh Lai, Advantest America, Inc., Product Marketing Manager

## Introduction

During the next four years, the NAND Flash market will have the third-highest rate of revenue growth among all semiconductor segments and demonstrate stronger bit growth than DRAMs, according to market research firm IC Insights. The surge in eMMC volume is being driven by mobile phones and tablets, which are demanding faster speeds and spurring the transition to UFS and other higher speed interfaces. Other applications such as Ultrabooks and enterprise-storage solutions are pushing SSD production higher while also requiring greater quality and faster speeds for both ONFi and Toggle NAND interfaces.

NAND technology is rapidly evolving as all major NAND manufacturers have begun migrating to sub-20-nm processes, effectively shrinking the basic 2-D NAND memory cell. In addition, chip makers are increasing the densities of Flash cells from two bits to three bits. Samsung Electronics has announced that it is producing 3-bits-per-cell NAND Flash memory devices using 10-nm process technology. Because many industry watchers believe that 2-D NAND scaling will reach its limit at the 10-nm node, the industry is moving to novel 3-D NAND lithography processes such as Samsung TCAT, Toshiba BiCS, SK Hynix DC-SF and Macronix BE-SONOS. These approaches are bringing dramatic changes to basic memory cell design.

Advances in NAND Flash technology have raised reliability and quality challenges for manufacturers, renewing the need for improved test coverage without raising the cost of test and dampening market growth. To achieve the required economic performance, any viable test solution must have an architecture capable of efficiently increasing throughout and yield.

## The Need for a Dedicated Test Solution

As with other commodity products, the per-bit ASP for NAND Flash has been dropping dramatically. According to Gartner Research, the ASP per gigabyte of NAND Flash declined from $7,870 in 1997 to just $0.25 by 2012. Although the per-bit ASP has stabilized a bit during the industry's recent consolidation, Gartner's forecast for the second quarter of 2013 calls for a declining CAGR of -18.7% for the period 2012-2017. As a result, NAND Flash manufacturers are facing great market pressure to reduce cost through measures such as minimizing the cost of

test. This has increased demand for highly economical NAND test solutions capable of increased throughput and yield.

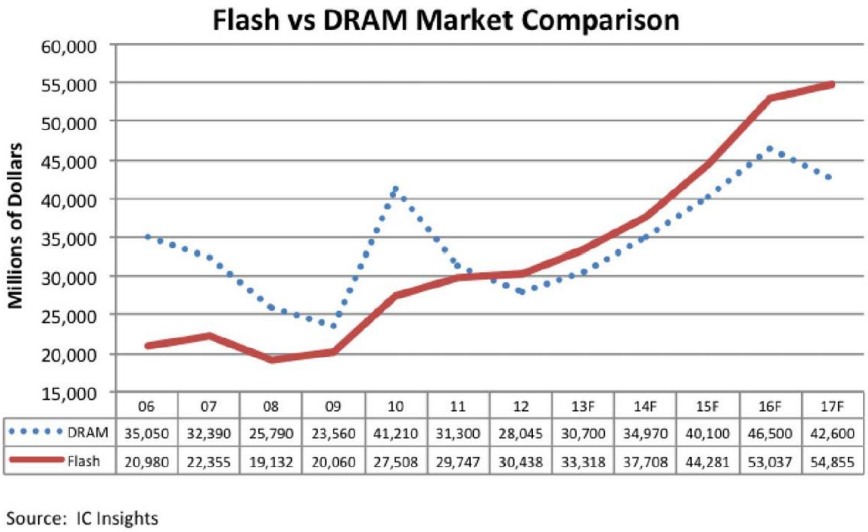
Because NAND Flash requires a set of unique test capabilities not commonly used for other types of memory ICs, a dedicated test solution is needed. Most importantly, the solution needs to be optimized for NAND Flash testing by stripping away any extraneous capabilities that would add to a tester's cost. Table 1 lists 10 key NAND Flash test functions and benefits, some of which will be covered in more detail later.

| ID | NAND Flash Test Functions | Benefits |
|---|---|---|
| 1 | Tester-per-site architecture — architecture in which there are many test sites in a system, each with its own independent test resources, such as test processor, algorithmic pattern generator, parametric measurement units, buffer memory, fail memory and others. | Enables fastest test time for NAND Flash because each test site only handles a small number of DUTs, independent of other test sites in the system. Although more test electronics increases cost, this test architecture still lowers the overall cost of test by delivering much faster test time. |
| 2 | Site-chain support — enables two or more test sites to be combined into a single test site with higher pin count. | Provides flexibility to test various NAND Flash interfaces, from low-pin-count devices such as eMMC to high-pin-count devices such as ONFi/Toggle NAND with one, two or more ports. |
| 3 | AC performance — accuracy and speed for next-generation ONFi/Toggle NAND and managed NAND such as eMMC. A tester optimized for NAND Flash needs to support at-speed testing up to 800 Mbps $\pm$ 10% with overall timing accuracy of < 200 ps to maximize yield. | Enables at-speed testing to guarantee quality while maximizing yield with high accuracy at low cost. Higher AC performance directly results in higher tester cost; therefore, it is necessary to optimize AC performance of the tester specifically for NAND Flash testing to minimize cost. |
| 4 | Real-time source synchronous function — capable of automatic, cycle-by-cycle adjustments to compensate for timing drift and jitter of critical timing parameters such as tDQSRE and tDQSQ. | Enables maximum yield and faster test time, especially when test speeds are higher than 400 Mbps. |
| 5 | Scalable, high-current PPS architecture — allows for flexible and high PPS pin-count per DUT. | Enables faster test time through concurrent, multi-die operation, such as NAND Flash array program and erase. |
| 6 | Flexible ECC analysis function — capable of on-the-fly analysis and ECC grading. | Enables maximum yield, flexibility and faster test time. |
| 7 | Bad-block management — allows for easy management of bad blocks in NAND devices, including the ability to mask bad blocks from being tested. | Enables faster test time and simpler test program generation for managing bad blocks. |
| 8 | Fail memory optimized for NAND — NAND Flash package is already approaching 1 terabit and increasing. It is not cost effective to have a large | Enables flexibility at lower cost. |

| | | |
|---|---|---|
| | fail memory in the tester to store fail bitmaps of every DUT. Additionally, it is even more costly when error logging is required during at-speed tests. Therefore, it is necessary to have a new fail memory design that is capable of providing the necessary functions as well as lowering cost. | |
| 9 | Buffer memory optimized for NAND — as with fail memory, it is not cost effective to have a large buffer memory to store custom data pattern to be used for programming and verifying the NAND Flash array, especially when it has to support at-speed tests. Therefore, it is also necessary to have a new buffer memory design that is capable of providing the necessary functions as well as lowering cost. | Enables flexibility at a lower cost. |
| 10 | eMMC test support — capable of supporting eMMC protocol and simple multi-DUT support. | Enables flexibility and simpler test program generation for faster TTM. |

**Table 1: Key test features for NAND Flash**

The NAND market surpassed the DRAM market in revenue for the first time in 2012, as shown in figure 1, and the gap is expected to grow for the foreseeable future. This market growth justifies the development of a dedicated NAND Flash test solution needed to accommodate evolving technologies, increase test coverage, reduce the cost of test and lower ASPs.




| | 06 | 07 | 08 | 09 | 10 | 11 | 12 | 13F | 14F | 15F | 16F | 17F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DRAM | 35,050 | 32,390 | 25,790 | 23,560 | 41,210 | 31,300 | 28,045 | 30,700 | 34,970 | 40,100 | 46,500 | 42,600 |
| Flash | 20,980 | 22,355 | 19,132 | 20,060 | 27,508 | 29,747 | 30,438 | 33,318 | 37,708 | 44,281 | 53,037 | 54,855 |

**Figure 1: Flash and DRAM market comparison**

## Choosing the Right Tester Architecture

The advantages of a tester-per-site architecture over shared-resource designs for testing NAND Flash are well documented, but it is worthwhile to examine some key differences among tester-per-site architectures to determine which are best suited for NAND Flash.

A true tester-per-site architecture that tests one DUT per site provides the highest throughput for NAND Flash. This architecture uses many test sites to achieve the required parallelism with each test site drawing upon its own independent test resources, such as test processor, algorithmic pattern generator, parametric measurement units, buffer memory, fail memory and others. However, the tester design is very expensive and may not offer the lowest overall cost of test. To achieve the optimal balance of system cost and throughput, this type of tester typically requires a workload of two to eight DUTs per site. This is evident by comparing cycle times and costs for two types of tester-per-site architectures, as shown in figure 2. Tester A can accommodate fewer DUTs per test site than tester B, and tester A offers a lower test time and lower cost of test.

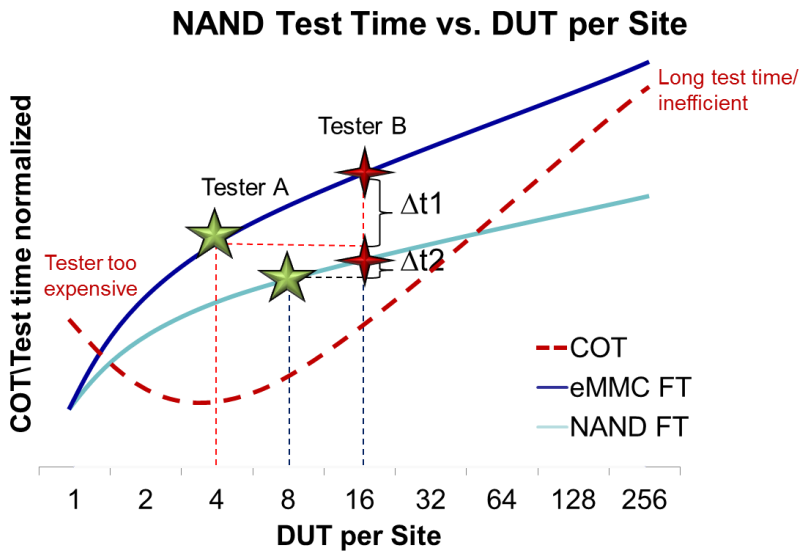



**Figure 2: Comparison of various tester-per-site architectures**

Normalized test times can differ with the test flow, as shown by the NAND FT and eMMC FT curves in figure 2. For a typical tester-per-site architecture, test time increases as a test site handles more DUTs. The rate of test time increase is highest between one to four DUTs per site, after which it gradually decreases. This performance is typical for NAND Flash test flows for several reasons:

- Test time for NAND Flash operation, especially array program, varies per Flash cell per DUT. The duration of test times can change with process differences.
- Per-DUT, unique data generation is required for test items such as trimming references and marking bad blocks in the array. When combined with testing multiple DUTs per site, it is typical for a tester to serially test each DUT in the test site, which results in longer test times. This bottleneck can be reduced by using a tester design that allows parallel testing of all DUTs in the test site, thereby reducing overall test time.
- Sharing some of the test resources in a test site among DUTs is a proven way to reduce tester costs. Test resources such as ADC and PMU, required for measuring voltages and currents on DUT pins, are typically shared among DUTs per test site. Applying these resources serially per-DUT can result in slightly longer test times, but will reduce hardware costs.
- The test processor is another test resource that is typically shared to save costs. It is responsible for running test programs that can include intensive data processing for every DUT in the test site. For instance, if a tester's hardware does not support real-time source-synchronous testing or on-the-fly ECC analysis, the test processor must perform significant post-processing of data for every DUT in the test site, which lengthens test time. However, when equipped with the proper hardware, all post-processing of data can be eliminated, resulting in shorter test time.
- eMMC testing requires protocol support based on a series of command-response exchanges between the tester and DUTs. However, response times can vary with each DUT by anywhere from two to 64 clock cycles, according to specifications. In addition to testing several DUTs per test site, which is typical for low-pin-count eMMC devices, the tester must conduct parallel polling of all DUTs in the test site for responses. If one DUT is ready to respond before the others, it must wait until all DUTs can respond in parallel. So eMMC test speeds are limited by the DUTs with the slowest response times.

## Real-time Source-Synchronous Testing

To satisfy market demand for NAND Flash devices with higher data-transfer rates, the industry has introduced DDR NAND Flash and managed NAND Flash such as eMMC. Already capable of supporting 200 MT/s and 400 MT/s data-transfer rates, these devices will need to accommodate even faster rates in the future. As data-transfer speeds increase, the NAND interface is upgraded with a source-synchronous capability by adding a bidirectional DQS signal to improve control of the data interface timing. This is another test requirement that can be best addressed by selecting the right tester architecture to improve yield and throughput.

There are two major DDR NAND interface designs. The Toggle Mode design is developed and supported by Toshiba and Samsung while the ONFi design is developed by the Open NAND Flash Interface Working Group and is supported by NAND manufacturers including SK Hynix, Micron, Intel and SanDisk. To achieve interoperability for these two designs, JEDEC and ONFi have developed and published the JESD230 NAND Flash Interface Interoperability Standard, which focuses on NAND Flash packages.

Both ONFi and Toggle Mode DDR NAND Flash have added bidirectional data strobes at data-transfer speeds of 133 MT/s and higher, as shown in tables 2 and 3. For next-generation devices with data-transfer speeds higher than 400 MT/s, source-synchronous testing becomes vital because timing margins are significantly reduced.

| Feature | ONFi 1.0 | ONFi 2.x | ONFi 3.0 | *ONFi NG (projection)* |
|---|---|---|---|---|
| Interface | SDR | NV-DDR | NV-DDR2 | *NV-DDRx?* |
| Maximum transfer speed | 50 MT/s | 200 MT/s | 400 MT/s | *533 MT/s ~ 800 MT/s?* |
| Data strobe | No | Yes | Yes | *Yes?* |

**Table 2: ONFi NAND Flash interface and projected roadmap (Source: JEDEC and independent projection)**

| Feature | Legacy | Toggle 1 | Toggle 2 | *Toggle NG (projection)* |
|---|---|---|---|---|
| Interface | SDR | DDR | DDR | *DDR?* |
| Maximum transfer speed | 40 MT/s | 133 MT/s | 400 MT/s | *533 MT/s ~ 800 MT/s?* |
| Data strobe | No | Yes | Yes | *Yes?* |

**Table 3: Toggle Mode NAND Flash interface and projected roadmap (Source: JEDEC and independent projection)**

Device data output presents a key challenge in supporting source-synchronous functionality. According to ONFi 3.0 specifications shown in figure 3 and table 4, there are two critical timing parameters – tDQSRE and tDQSQ – which have very wide ranges and are extremely sensitive to process-voltage-temperature (PVT). Since most of today's testers do not have real-time source-synchronous support, users often rely on a process called "training" to determine these parameters. Training is a time-consuming operation in which a simplified read pattern is looped while strobe-timing edges are finely adjusted over a wide range until the desired result is found.

Because tDQSRE and tDQSQ timing parameters are sensitive to PVT, training must be performed for every DUT whenever voltage conditions are modified in the test flow. This can involve significant test-time overhead. In addition, device operation may result in temperature changes. In combination with jitter, this can affect timing parameters during testing, which cannot be resolved with training. If a tester does not have a source-synchronous function to automatically perform real-time compensation, then yield can be lost, particularly at data-transfer rates higher than 400 MT/s where data eye is significantly reduced.

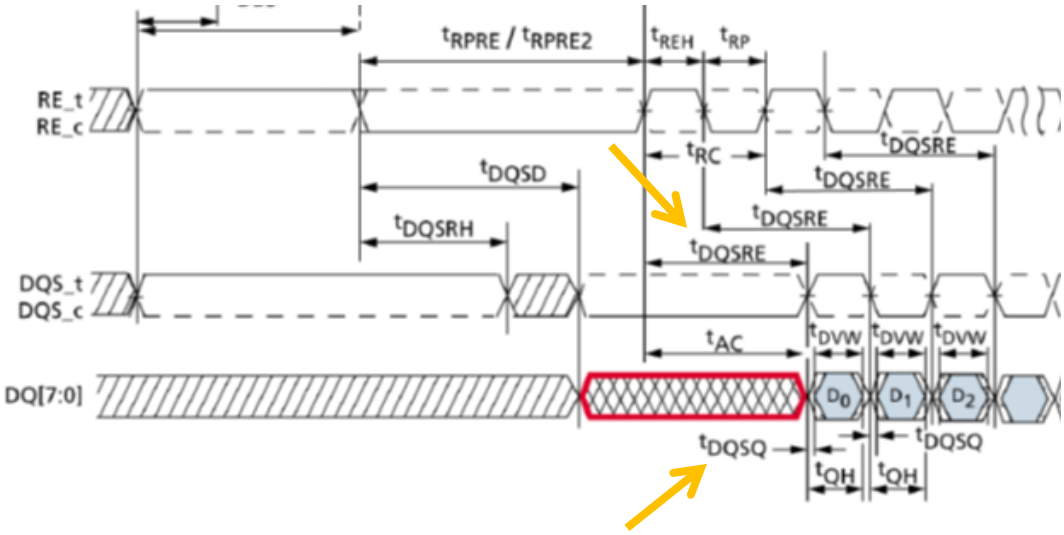



Figure 3: NV-DDR2 data output cycle timing (Source: JEDEC ONFi 3.0 specification)

| | Condition | Min. | Max. | Unit |
|---|---|---|---|---|
| tDQSRE | | 3 | 25 | ns |
| tDQSQ | ~100 MHz (tRC = 10 ns) | -- | 0.8 | ns |
| | ~133 MHz (tRC = 7.5 ns) | -- | 0.6 | ns |
| | ~166 MHz (tRC = 6 ns) | -- | 0.5 | ns |
| | ~200 MHz (tRC = 5 ns) | -- | 0.4 | ns |

Table 4: tDQSRE and tDQSQ specifications (Source: JEDEC ONFi 3.0 specifications)

As illustrated in figure 4, a tester with real-time source-synchronous capability can automatically compensate for the tDQSRE and tDQSQ dispersions over multiple devices on a cycle-by-cycle basis, despite changing PVT. This enables testing with guaranteed data eye to maximize yield when running at data rates higher than 400 MT/s. This tester also is able to eliminate all related training operations, further reducing test time.

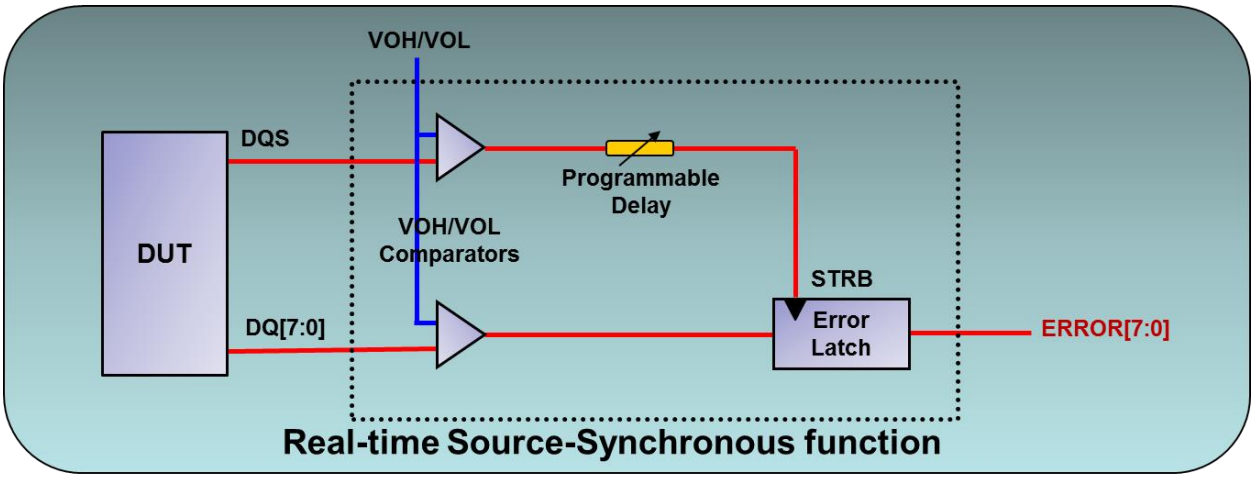



**Figure 4: Real-time source-synchronous function**

Data-transfer rates of 400 MT/s and higher also are being used in managed NAND Flash, such as eMMC typically used for mobile applications. As with ONFi and Toggle NAND interfaces, the eMMC interface is moving to DDR and adding source-synchronous capability to better support higher data-transfer rates. The real-time source-synchronous function also can support eMMC 5.0 requirements, as shown in table 5.

| | eMMC4.41 | eMMC4.51 | eMMC5.0 |
|---|---|---|---|
| Clock frequency | 0~52 MHz | 0~200 MHz | 0~200 MHz |
| Maximum bandwidth | 104 MB/s | 200 MB/s | 400 MB/s |
| Interface | SDR-104 | SDR-200 | DDR-200 |
| Data strobe | No | No | Yes |

Table 5: eMMC interface and projected roadmap (Source: JEDEC and independent projection)

## Saving Time with On-the-Fly ECC Analysis

As NAND manufacturers have advanced to sub-20-nm process lithography, they have been able to increase the number of bits per cell to meet growing market demand for higher storage densities. Using 10-nm process technology, a floating-gate NAND memory cell holds as few as 10 electrons, according to the trend. At 3 bits per cell, the data stored in such memory cell changes by adding or removing just one or two electrons in the floating gate. The adoption of novel 3-D lithography processes opens the door to further advances. All of these developments improve the quality and reliability of NAND Flash to help it meet the performance demands for SSDs in Ultrabooks and enterprise-storage solutions. As a result, NAND manufacturers are more dependent upon increasing ECC capabilities, which can be addressed by the tester architecture.

The majority of testers on today's market do not offer real-time ECC support. This puts the burden on users to perform a post-processing step to determine ECC solutions for every DUT, a time-consuming process, especially for test flows that perform many read operations with ECC analysis enabled. In addition, it requires substantial tester resources including a large error-capture RAM to store fail bitmap of every DUT as well as high-performance test processors to analyze those fail bitmaps and develop ECC solutions. When done, this results in a significant increase in cost of test to support ECC analysis.

Developing a low-cost ECC analysis function with on-the-fly analysis requires the following attributes:

- On-the-fly ECC analysis: To eliminate all related test-time overhead, ECC analysis must be done on-the-fly across all DUTs. As a side benefit, this also reduces tester cost by not requiring costly tester resources such as a large error-capture RAM and many high-performance test processors.
- Flexibility: NAND is used by a wide variety of customers, each with its own proprietary controller and ECC algorithm supporting different data organization (figure 6). Any test solution must have the flexibility to support multiple ECC algorithms.
- Concurrent execution of multiple ECC algorithms: To save test time, testers also must be able to run several ECC algorithms simultaneously.
- Support of ECC grading: In addition to reporting pass/fail status, a practical test solution must have the capability to store fail counts per ECC sector so that the data can be analyzed, graded and binned as necessary. This gives NAND manufacturers the flexibility to sell their devices to customers that require different levels of fail-bits per ECC sector, thus, allowing manufacturers to increase profitability by providing the right products to the right customers while also reducing the volume of rejected devices.

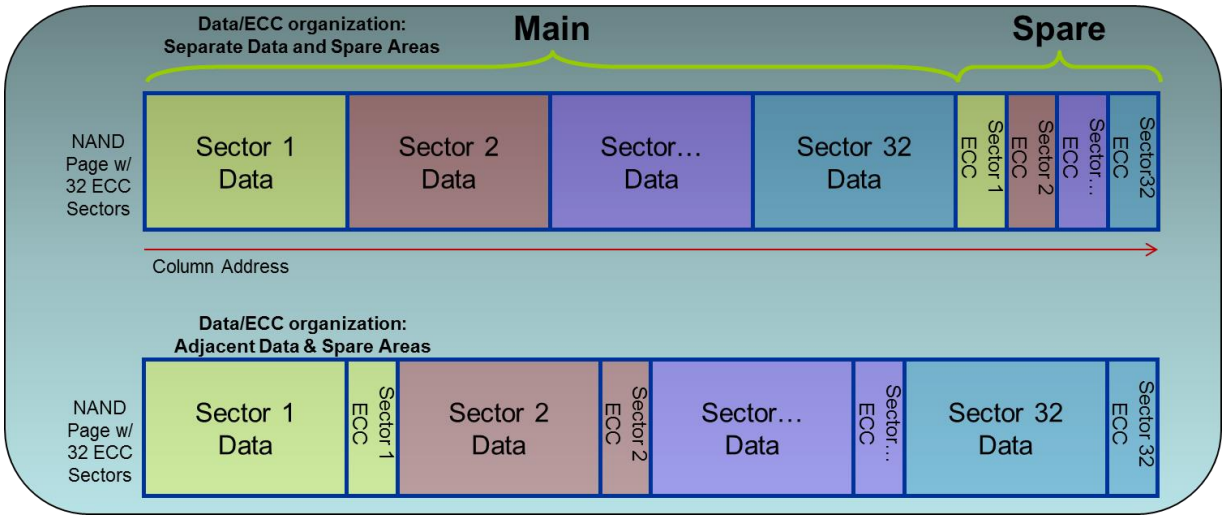



**Figure 5: Data/ECC organizations**

## Boosting Throughput with High-Current PPS

To achieve the higher densities and integration needed for mobile applications, NAND manufacturers are stacking as many as 16 dies per package. Very long cycle times will be required to test all the bits in a package with a total capacity of 1 terabit or more, and this slow throughput will directly increase the cost of test. One solution being pursued is to concurrently test all dies in a package. While this approach is especially promising for long test time items such as array program and erase, it requires a significant amount of power supply current per DUT. Testers with scalable power supply architectures enable flexible PPS pin-count per DUT, which can reduce test times as well as cost of test.

If done sequentially per die, program and erase operations for typical NAND Flash arrays consume about 100 mA of ICC. To reduce test time, these operations can be performed concurrently on all dies in the package. The ICC current required is directly proportional to the number of dies in the operation. The impact of test time for array program/erase operations varies among testers with different PPS pin-count per DUT. The reduction in test time is significant for testers with high PPS pin-count per DUT, especially in packages with more stacked dies.

| Parameter | # of die in parallel | Max. | Unit |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Array program/erase current (ICC at VCC = 3.7 V) | 1 | 100 | mA |
| | 2 | 200 | mA |
| | 4 | 400 | mA |
| | 8 | 800 | mA |
| | 12 | 1,200 | mA |
| | 16 | 1,600 | mA |

**Table 6: Example of array program and erase current requirement**

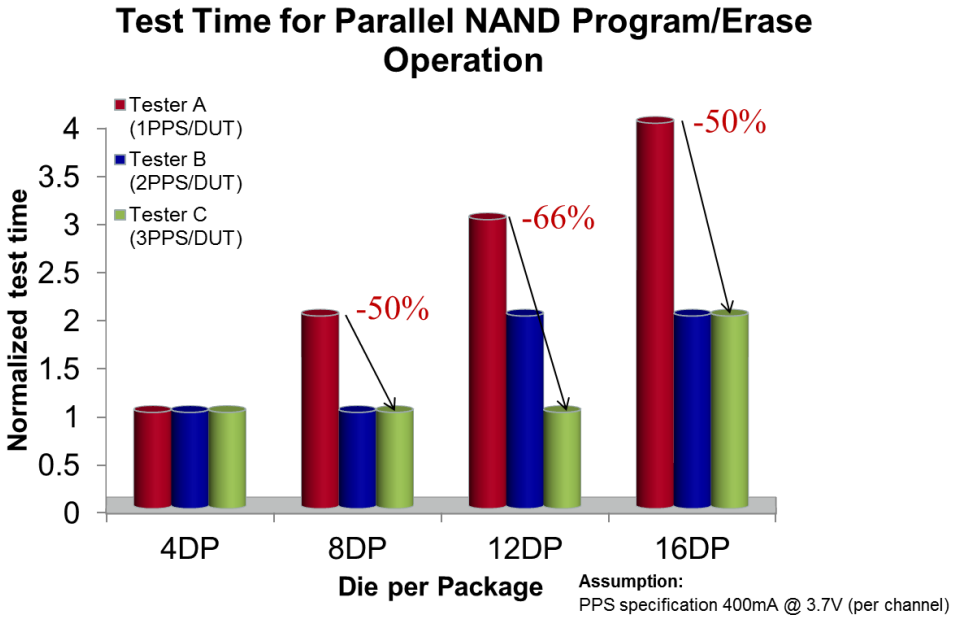



**Figure 6: Test time for parallel NAND program/erase operations versus number of PPS per DUT**

Many of today's testers do not have PPS resources with the needed voltage and ICC current to support concurrent array program/erase operation for up to 16 dies per package, although some testers allow ganging of two or more PPS channels to boost the ICC current. Unfortunately, most existing testers have a fixed PPS pin-count. This means that either test time must be extended to allow for serialization of the array program and erase operations for all dies in the package or parallelism must be reduced to handle more PPS channels per DUT. Whichever approach is used, the result is lower throughput and a higher cost of test.

However, a tester with a scalable PPS architecture can accommodate a range of PPS channels per DUT. This enables faster testing and a significantly lower cost of test.

## Summary

Driven by the booming markets for mobile phones, tablets and SSDs, the NAND Flash market is projected to continue growing for the foreseeable future. Faced with market demand for greater device densities and performance along with higher product quality and reliability, NAND manufacturers are developing innovative, new NAND Flash technologies. In turn, this increases demand for greater test coverage, which drives up the cost of test. To manage escalating costs, NAND manufacturers need a dedicated, economical NAND Flash test solution. It is clear that the solution requires a scalable tester-per-site architecture to deliver optimized AC performance and other NAND-specific capabilities that will increase both throughput and yield.