swissbit®

Application Note

AN2108de

Comparing Specifications

[1]Dieses Werk steht unter der Creative-Commons-Lizenz „Namensnennung 4.0 International“. Um eine Kopie dieser Lizenz zu sehen, besuchen Sie http://creativecommons.org/licenses/by/4.0/

swissbit®

Inhaltsverzeichnis

1 Einleitung

Ob ein Flash-Speichermedium prinzipiell für den beabsichtigten Einsatzzweck geeignet ist, lässt sich vorab anhand des Datenblatts ermitteln. Neben dem Betriebstemperaturbereich sind für die meisten Anwendungen hauptsächlich die Lebensdauer und die Geschwindigkeit maßgeblich. Hierzu listen die Hersteller im Datenblatt Werte meist anhand einer Auswahl üblicher Testmethoden. Je nach Hersteller kommen jedoch unterschiedliche Testmethoden zur Anwendung, wodurch die resultierenden Werte nicht immer leicht mit Produkten anderer Hersteller vergleichbar sind.

Die typischen Methoden, Programme und Begriffe, die für solche Messungen verwendet werden, werden im Folgenden näher erläutert, um die Charakteristika verschiedener SSDs besser vergleichen und mit dem Einsatzzweck abgleichen zu können.

2 Geschwindigkeit

Für Geschwindigkeitsmessungen ist es maßgeblich, mit welchem Zugriffsmuster gelesen und geschrieben wird. Zudem hat das Verteilungsmuster der vorangegangenen Schreibzugriffe einen sehr großen Einfluss auf die nachfolgenden Lese- und Schreibgeschwindigkeiten.

Grundsätzlich sollten alle Speichermedien vor einer Geschwindigkeitsmessung eine Vorbehandlung erfahren, bei der alle Speicheradressen mindestens einmal beschrieben werden. Da es bei NAND-Flash keine feste Zuordnung gibt zwischen den physikalischen Flash-Adressen und den logischen Adressen, die das Betriebssystem sieht, werden die aktuellen Zuordnungen in Verwaltungstabellen nachgeführt, die ebenfalls im Flash gespeichert sind. Durch die Vorbehandlung sind allen logischen Adressen auch physikalische Adressen im Flash zugeordnet. Bei einem Lese- oder Schreibzugriff auf eine logische Adresse muss somit zuerst die dazugehörige physikalische Adresse gesucht bzw. nach einem Schreibzugriff die alte Adresse invalidiert werden. Fehlt die Vorbehandlung, ist der erste Schreibzugriff auf die logische Adresse schneller als die nachfolgenden Zugriffe. Entsprechend sollte im Datenblatt vermerkt sein, falls der Schreibtest auf ein unbenutztes Medium erfolgte.

2.1 Sequentieller Zugriff

Das einfachste Zugriffsmuster bei Geschwindigkeitstests ist der sequentielle Zugriff. Dabei wird das Medium über den gesamten Adressraum aufsteigend mit Daten gefüllt und die Daten in großen Segmenten von mindestens 128 KiB vom Host übertragen. Die Daten selbst bestehen aus Zufallsdaten, also Daten mit maximaler Entropie, um das Messergebnis bei (seltenen) Speichermedien mit integrierter Kompression nicht zu verfälschen. Auch bei Medien ohne Kompression sollten Daten, die nur aus binären Nullen bestehen, vermieden werden, da einige Flash-Controller Nullen erkennen, und statt die Nullen auf den Flash zu schreiben, einfach die Einträge dieser Speicheradressen in den Verwaltungstabellen löschen, wodurch bei einem Lesezugriff auch nur wieder Nullen zurückgeliefert werden, ohne aus dem Flash lesen zu müssen.

Der Wert unter *Sequential Write* im Datenblatt ist daher die durchschnittliche Schreib-

geschwindigkeit über die gesamte Kapazität, oder – wenn entsprechend vermerkt – nur von einem Ausschnitt der Gesamtkapazität bzw. der Maximalwert statt des Durchschnittswertes. Hier tritt das Problem auf, dass sich der arithmetische Mittelwert über die gesamte Kapazität nicht mit reduzierten Bereichen oder Maximalwerten vergleichen lässt, da heute viele Medien über einen schnellen pSLC-Cache verfügen, der über mehrere Prozent der Gesamtkapazität Schreibzugriffe aufnehmen kann, bevor auf den langsameren TLC oder QLC gewechselt werden muss.

Die sequentielle Schreibgeschwindigkeit erreicht den höchsten Wert unter allen Schreibtests, da der interne Verwaltungsaufwand minimal ist und die Garbage-Collection nicht beansprucht wird. Dieser Wert stellt daher eine gute Orientierung für Anwendungsfälle dar, bei denen überwiegend große Dateien wie Fotos oder Videos geschrieben werden.

Entsprechend zum sequentiellen Schreibzugriff gibt es auch einen Wert für den sequentiellen Lesezugriff (*Sequential Read*). Zuvor wurde das Medium über den gesamten Adressraum aufsteigend gefüllt. Dadurch ist auch die physikalische Verteilung im Flash sequentiell, wodurch die Suche nach den Adressen in den Tabellen einfach ist und die Lesezugriffe dadurch schnell sind, zumal auch das Lesen effizient in großen Datensegmenten erfolgt.

Auch hier lassen sich Durchschnittswerte über das ganze Laufwerk nicht mit Ausschnitten oder Maximalwerten vergleichen, da ein evtl. vorhandener pSLC-Cache noch Daten vom vorangegangenen Schreibzugriff enthält und die Lesezugriffe beschleunigt.

2.2 Zufälliger Zugriff

Bei Anwendungsfällen mit vielen kleinen Dateien oder Datenbanken sowie bei Mehrbenutzer-Systemen oder generell bei Servern sind die Zugriffe auf das Speichermedium überwiegend zufällig verteilt. Zudem sind die gelesenen und geschriebenen Datenmengen eher klein. Für solche Anwendungsfälle orientiert man sich besser an den Werten für *Random Write* und *Random Read*. Sofern nicht anders angegeben, beziehen sich diese Werte auf Datensegmente von 4 KiB, die zufällig über das ganze Medium geschrieben und gelesen werden. Dabei wird jedes 4 KiB Segment genau einmal beschrieben bzw. gelesen. Anschließend wird wieder der Durchschnittswert für das ganze Medium angegeben. Wie bei jedem Geschwindigkeitstest wurde das gesamte Medium zuvor komplett sequentiell beschrieben, um eine bekannte und wiederholbare Startbedingung zu haben.

Auch hier wird ein Medium beim Schreiben und Lesen besser abschneiden, wenn nicht auf 100 % des Mediums zugegriffen oder der Maximalwert statt des arithmetischen Mittels angegeben wird. Wählt der Hersteller bei diesem Test eine Segmentgröße von mehr als 4 KiB, beschleunigt dies die Zugriffe ebenfalls. Die Zugriffsgröße von 4 KiB wurde gewählt, da dies die kleinste Standardblockgröße von Dateisystem ist, also die kleinste Einheit die gelesen oder geschrieben werden kann. Server-Anwendungen erzeugen daher hauptsächlich diese Zugriffsgröße.

Die Geschwindigkeitsangaben für zufällige Zugriffe erfolgen meist in der Einheit *Input/-Output operations Per Second* kurz *IOPs*. Dies ist die Anzahl der übertragenen 4 KiB Segmente pro Sekunde. Die Multiplikation dieses Wertes mit 4 KiB ergibt dann die Geschwindigkeit in KiB/s.

Zufällige 4 KiB Zugriffe erreichen nur eine geringe Geschwindigkeit. Zum einen steigt der Verwaltungsaufwand aufgrund der kleinen Segmentgröße auf der gesamten Strecke zwischen Host und dem NAND-Flash stark an, zum anderen ist während des Schreibens die Garbage-Collection[2] unter Volllast und beim Lesen ist die Suche nach der physikalischen Adresse in den Verwaltungstabellen zeitraubend.

Dieser Test – sofern mit 4 KiB über die gesamte Kapazität ausgeführt – stellt für Lese- und Schreibzugriffe die absolute Geschwindigkeits-Untergrenze dar. Kein Zugriffsmuster ist langsamer. Somit zeigt dieser Test die Minimalgeschwindigkeit des Mediums.

[2] Zu Garbage-Collection siehe Application Note AN2101

2.3 CrystalDiskMark

Sehr häufig findet man heute Geschwindigkeitsangaben zu sequentiellen und zufälligen Zugriffen, die auf Messungen mit CrystalDiskMark beruhen. Dieses nur für Windows verfügbare Testprogramm hat mehrere Vorteile. Es ist kostenlos, einfach zu bedienen, benötigt nur wenige Minuten Laufzeit und ist nicht destruktiv, da es dateibasiert ist und keine bestehenden Daten überschreibt. Die Geschwindigkeitsmessung mit CrystalDiskMark kann daher mit vorhandenen Medien einfach selbst durchgeführt und mit Datenblättern verglichen werden. Dabei ist zu beachten, dass nur Werte verglichen werden können, die mit identischer Hauptversionsnummer von CrystalDiskMark gemessen worden sind.

Der große Nachteil von CrystalDiskMark ist jedoch die kurze Laufzeit. Die maximale Dateigröße ist begrenzt und in der Standardeinstellung nur 1 GiB. Speicherarchitekturen, die über einen pSLC-Cache oder bei Enterprise-SSDs über einen großen DRAM verfügen, erreichen somit sehr gute Werte, da der massive Einbruch der Geschwindigkeit erst bei größeren Datenmengen erfolgt, wenn der Wechsel auf TLC/QLC erfolgt.

2.4 Burst Mode

Der *Burst Mode* oder auch *Burst Transfer Mode* bezeichnet die maximale Geschwindigkeit, die das Medium erreichen kann. Diese tritt auf, wenn beim Schreiben die Daten im Puffer des Flash-Controllers aufgenommen werden bzw. beim Lesen aus diesem Puffer an den Host geschickt werden. Zugriffe auf den Flash finden nicht statt. Einen ausreichend schnellen Flash-Controller vorausgesetzt, entspricht diese Geschwindigkeitsangabe der Übertragungsgeschwindigkeit der Schnittstelle zum Speichermedium, also z. B. 600 MB/s bei SATA III. Eine Praxisrelevanz hat dieser Wert nicht, da er nur für Sekundenbruchteile erreicht wird.

3 Lebensdauer

Die Lebensdauer eines Flash-Speichermediums definiert sich über die Anzahl der Lösch-Zyklen pro Flashblock, nach denen das stromlose Medium die Daten noch eine definierte Zeitspanne bei einer definierten Temperatur halten kann. Eine typische Kombination sind 3000 Löschzyklen pro Block, nach denen der Flash die Daten bei +40 °C noch ein Jahr halten kann. Die Firmware des Speichermediums sorgt dafür, dass alle Flashblöcke gleichmäßig altern.

Wie schon bei der Geschwindigkeit ist für die Lebensdauer ebenfalls das Zugriffsmuster beim Schreiben entscheidend. Für die Messungen mit den verschiedenen Methoden findet auch hier wieder die Vorbehandlung mit dem kompletten sequentiellen Beschreiben des Mediums statt. Dann werden die Löschzähler des Mediums ausgelesen und anschließend das Testprogramm gestartet. Dabei wird die Datenmenge gemessen, die vom Host an das Speichermedium transferiert wird. Aus dieser übertragenen Datenmenge und der Zunahme der Lösch-Zyklen lässt sich dann die über die gesamte Lebensdauer erreichbare Datenmenge berechnen. Dieser Wert findet sich als *Terabytes Written* (*TBW*) im Datenblatt.

Oft sind auch die *Drive Writes per Day* (*DWPD*) angegeben. Hierzu wird der TBW-Wert durch die Kapazität des Mediums geteilt und anschließend noch durch eine definierte Anzahl an Tagen.

Beispiel: Ein 512 GB Medium erreicht bei einem bestimmten Test 560 TBW bis der Datenerhalt im stromlosen Zustand auf ein Jahr gefallen ist.

$$\frac{560\,\mathrm{TBW}}{512\,\mathrm{GB}\cdot 3\,\mathrm{Jahre}} = 1\,\mathrm{DWPD}$$

Über eine angenommene Einsatzdauer von drei Jahren ergibt sich ein DWPD. Jeden Tag, für die Dauer von drei Jahren, kann also mit diesem Test einmal die Datenmenge, die der Speicherkapazität von 512GB entspricht, geschrieben werden.

3.1 Sequentielles Schreiben

Beim sequentiellen Schreiben wird das Medium wiederholt von Anfang bis Ende mit großen Datensegmenten gefüllt. Der Garbage-Collector muss nicht tätig werden, da automatisch wieder Flashblöcke frei werden. Es müssen intern keine Daten verschoben werden, um neue leere Blöcke zu schaffen. Somit wird hier normalerweise das Maximum an Lebensdauer erreicht, da neben dem Schreiben der Nutzdaten kein weiterer nennenswerter Verschleiß auftritt. Allerdings kann es bei Architekturen mit pSLC-Cache dazu führen, dass durch den geringen Verschleiß des TLC/QLC-Bereichs der pSLC-Cache zuerst seine spezifizierte Lebensdauer erreicht und die Gesamtlebensdauer damit limitiert.

3.2 Zufälliges Schreiben

Beim zufälligen Schreiben ist der Garbage-Collector unter Volllast, da er ständig Nutzdaten verschieben muss, um wieder freie Flashblöcke zu schaffen. Intern muss ein vielfaches von der Datenmenge bewegt werden, die vom Host auf das Medium geschrieben wird. Sofern nichts anderes angegeben, ist die Segmentgröße wieder 4 KiB. Dieser Test stellt für die Lebensdauer den schlechtesten Fall dar und sollte für nahezu sämtliche Anwendungsszenarien unrealistisch sein.

3.3 JEDEC-Tests

Die *JEDEC Solid State Technology Association* hat Normen zur Standardisierung für Flash-Medien erstellt. Für die Lebensdauerangaben hat sie Testvorschriften in den Normen JESD218 und JESD219 herausgegeben. Darin werden für Server und Mehrbenutzersysteme („Enterprise“) und Einzelplatzrechner („Client“) Testvorschriften definiert, die typische Anwendungsfälle wiederspiegeln sollen. Die damit gewonnenen Messwerte stellen deutlich realistische Angaben zur Lebensdauer dar, als die beiden Extrema vom sequentiellen und zufälligen Schreiben. Lebensdauerangaben für Enterprise und Client sollten sich in jedem Datenblatt finden lassen.

3.3.1 Enterprise

Der sogenannte *Enterprise Workload* stellt typische Server-Anwendungen nach. Er besteht zu 2/3 aus 4 KiB Segmenten. Den Rest bilden verschiedene größere und kleinere Datensegmente. Auch Mehrbenutzersysteme zeigen in den Zugriffsmustern eine gewisse Lokalität, daher erfolgt die Hälfte der Zugriffe auf nur 5 % der Kapazität. Neben dieser Einschränkung erfolgen alle Zugriffe zufällig. Der Test läuft eine Woche.

Die JEDEC fordert, dass die geschriebenen Daten maximale Entropie haben müssen, also nicht komprimierbar sein dürfen für den Fall, dass das Speichermedium Kompression unterstützt.

Der Enterprise Workload profitiert nicht von einem pSLC-Cache. Aufgrund der vielen zufälligen Zugriffe erreicht er die besten Werte auf Architekturen mit kleinen internen Verwaltungsgrößen. Entsprechend führt eine Optimierung auf den Enterprise Workload zu einer Verschlechterung der sequentiellen Schreibgeschwindigkeit.

3.3.2 Client

Während der Enterprise Workload ein synthetischer Test ist, handelt es sich beim *Client Workload* um die Aufzeichnung der Schreibzugriffe in einem realen System. Diese Aufzeichnung wird mehrfach abgespielt und anschließend wird die Ausdauer des Mediums hochgerechnet. Im Gegensatz zum Enterprise-Workload, der nur aus Schreibbefehlen besteht, enthält der Client Workload zudem Trim- und Flush-Cache-Kommandos. Während die Trim-Kommandos den Garbage-Collector unterstützten und den Verschleiß reduzieren, erzwingen die Flush-Cache-Kommandos das Wegschreiben von Datenpaketen zum Flash, die noch im RAM des Flash-Controllers gehalten werden.

Die kleinste Einheit, die im Flash adressiert werden kann, ist die Page. Mehrere Pages sind

aber physikalisch oder logisch miteinander verbunden, wodurch pro Schreibzugriff mehrere hundert Kilobyte in einem Zug geschrieben werden müssen. Daher warten die Betriebssysteme bei einem Schreibbefehl mehrere Sekunden und sammeln in dieser Zeit weitere Daten, bevor sie die Daten zum Speichermedium weiter senden. Der Flash-Controller selbst wartet noch ein paar Millisekunden, um sicher zu gehen, dass er alle Daten vom Host erhalten hat, bevor er den Schreibvorgang auf den Flash startet. Die Flush-Cache-Kommandos verhindern dieses Vorgehen. Sie weisen das Betriebssystem und den Flash-Controller an, die bisher erhaltenen Daten sofort auf den Flash zu schreiben. Dadurch werden große Flash-Bereiche beschrieben, die dann nur wenige Nutzdaten enthalten.

Von Anwendungen und vom Dateisystem werden Cache-Flush-Kommandos verwendet, um die Daten vor einem unerwarteten Stromausfall zu schützen bzw. Fixpunkte für das Journal des Dateisystems zu erstellen.

Durch die vielen Flush-Cache-Kommandos im Client Workload profitieren wie beim Enterprise Workload Architekturen mit kleinen internen Verwaltungsgrößen. Da im Gegensatz zum Enterprise Workload beim Client Workload der verwendete Adressraum auf knapp 60 GB beschränkt ist, zeigen Systeme mit großem pSLC-Cache eine hohe Ausdauer, da Teile oder sogar der gesamte Test im Cache ablaufen kann und kein Transfer in den TLC/QLC mehr stattfindet.

4 Zusammenfassung

Geschwindigkeitswerte und Lebensdaueranga-ben lassen sich zwischen verschiedenen Speichermedien nur vergleichen, wenn exakt die gleiche Testmethodik angewendet wurde. Gehen die Rahmenbedingungen aus dem Datenblatt nicht hervor, kann nicht beurteilt werden, ob die angegebenen Werte vergleichbar sind.

Bei der Lebenserwartung geben die JEDEC-Workloads einen genauen Testablauf vor, so dass die Werte verschiedener Hersteller direkt vergleichbar sind. Nur muss dabei berücksichtigt werden, ob der eigene Anwendungsfall überhaupt dem Client bzw. dem Enterprise Workload entspricht. So entsprechen typische Industrieanwendungen zwar durchaus dem Schreibmuster des Client Workloads, jedoch kommen Flush-Cache-Kommandos meist nicht in dieser Häufigkeit vor, so dass eine wesentlich größere Lebensdauer zu erwarten ist.

Kontaktieren Sie uns

Hauptsitz	**Swissbit AG** Industriestraße 4 9552 Bronschhofen Schweiz	Tel. +41 71 913 03 03 sales@swissbit.com
Deutschland (Berlin)	**Swissbit Germany AG** Bitterfelder Straße 22 12681 Berlin Deutschland	Tel. +49 30 936 954 0 sales@swissbit.com
Deutschland (München)	**Swissbit Germany AG** Leuchtenbergring 3 81677 München Deutschland	Tel. +49 30 936 954 400 sales@swissbit.com
Nord- und Südamerika	**Swissbit NA Inc.** 238 Littleton Road, Suite 202B Westford, MA 01886 USA	Tel. +1 978-490-3252 salesna@swissbit.com
Japan	**Swissbit Japan Co., Ltd.** CONCIERIA Tower West 2F 6-20-7 Nishishinjuku Shinjuku City, Tokyo 160-0023 Japan	Tel. +81 3 6258 0521 sales-japan@swissbit.com
Taiwan	**Swissbit Taiwan** 3F., No. 501, Sec.2, Tiding Blvd. Neihu District, Taipei City 114 Taiwan, R.O.C.	Tel. +886 912 059 197 salesasia@swissbit.com
China	**Swissbit China**	Tel. +886 958 922 333 salesasia@swissbit.com

Disclaimer:
The information in this document is subject to change without notice. Swissbit AG ("SWISSBIT") assumes no responsibility for any errors or omissions that may appear in this document, and disclaims responsibility for any consequences resulting from the use of the information set forth herein. SWISSBIT makes no commitments to update or to keep current information contained in this document. The products listed in this document are not suitable for use in applications such as, but not limited to, aircraft control systems, aerospace equipment, submarine cables, nuclear reactor control systems and life support systems. Moreover, SWISSBIT does not recommend or approve the use of any of its products in life support devices or systems or in any application where failure could result in injury or death. If a customer wishes to use SWISSBIT products in applications not intended by SWISSBIT, said customer must contact an authorized SWISSBIT representative to determine SWISSBIT willingness to support a given application. The information set forth in this document does not convey any license under the copyrights, patent rights, trademarks or other intellectual property rights claimed and owned by SWISSBIT.

ALL PRODUCTS SOLD BY SWISSBIT ARE COVERED BY THE PROVISIONS APPEARING IN SWISSBIT'S TERMS AND CONDITIONS OF SALE ONLY, INCLUDING THE LIMITATIONS OF LIABILITY, WARRANTY AND INFRINGEMENT PROVISIONS. SWISSBIT MAKES NO WARRANTIES OF ANY KIND, EXPRESS, STATUTORY, IMPLIED OR OTHERWISE, REGARDING INFORMATION SET FORTH HEREIN OR REGARDING THE FREEDOM OF THE DESCRIBED PRODUCTS FROM INTELLECTUAL PROPERTY INFRINGEMENT, AND EXPRESSLY DISCLAIMS ANY SUCH WARRANTIES INCLUDING WITHOUT LIMITATION ANY EXPRESS, STATUTORY OR IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.