SSDs: Grundlagen + Konfiguration für hohe Performance

Werner Fischer, Technology Specialist Thomas-Krenn.AG

Thomas Krenn Roadshow 2010

11.10. Berlin
12.10. Hamburg
14.10. Köln
19.10. Frankfurt
20.10. Stuttgart
21.10. Zürich
05.11. Wien

Thomas-Krenn.AG®
Speed is (y)our success
TK SERVER

Agenda

1) **SSD Aufbau**

2) **Schreibtechniken**

3) **Anwendungsbeispiele**

4) **Konfigurationstipps**

Agenda

1) **SSD Aufbau**
   - **Speicherzellen**
   - **Pages**
   - **Blöcke**
   - **Planes**
   - **Dies**
   - **TSOPs**
   - **SSDs**

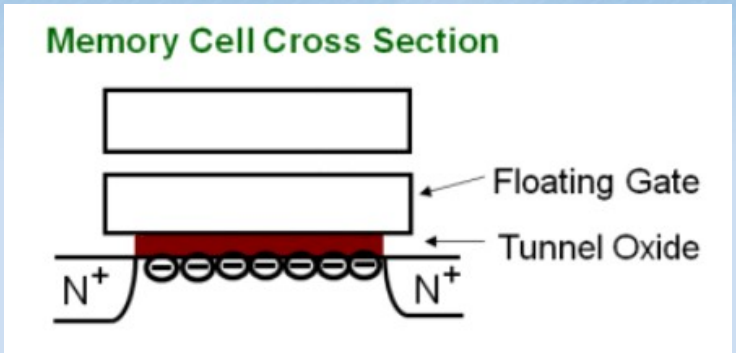
2) **Schreibtechniken**

3) **Anwendungsbeispiele**

4) **Konfigurationstipps**

1) Aufbau

- **Speicherzellen**
  - NAND Speicherzelle = MOS Transistor mit Floating Gate
  - speichert Ladung permanent

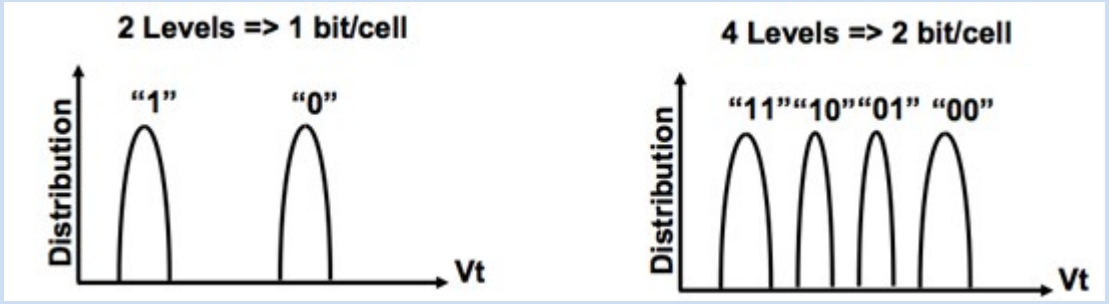


Quelle: Intel

  - bei der Programmierung werden Elektronen am Floating Gate abgelegt
  - ein Löschen entfernt die Elektronen
  - ein Program/Erase (p/e) Zyklus entspricht einem Round-Trip der Elektronen
  - jeder p/e Zyklus führt zu einer Abnützung des Tunnel Oxide
  - Lebensdauer dadurch begrenzt – wird in Anzahl der möglichen p/e Zyklen gemessen

1) Aufbau

- **Speicherzellen**
  - SLC (Single Level Cell) → 1 Bit pro Speicherzelle
  - MLC (Multi Level Cell) → 2 Bits pro Speicherzelle



Quelle: anandtech.com

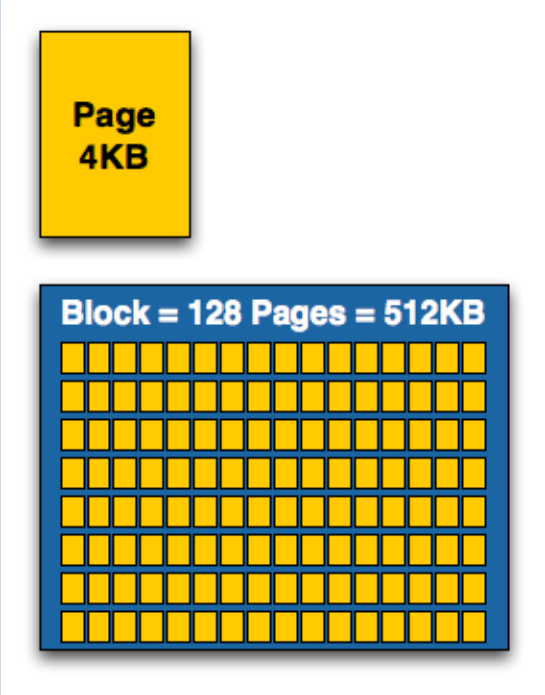
  - TLC (Triple Level Cell) → 3 Bits pro Speicherzelle

1) Aufbau

- **Pages: Zusammenfassung mehrerer Speicherzellen**
  - eine Page ist die kleinste Struktur einer SSD, die gelesen oder geschrieben werden kann
  - derzeit meist 4 KiB (4.096 Bytes) – entspricht bei MLC 16.384 Speicherzellen
  - bei den neuen 25nm Flash Chips von Intel/Micron 8 KiB (8.192 Bytes)

1) Aufbau

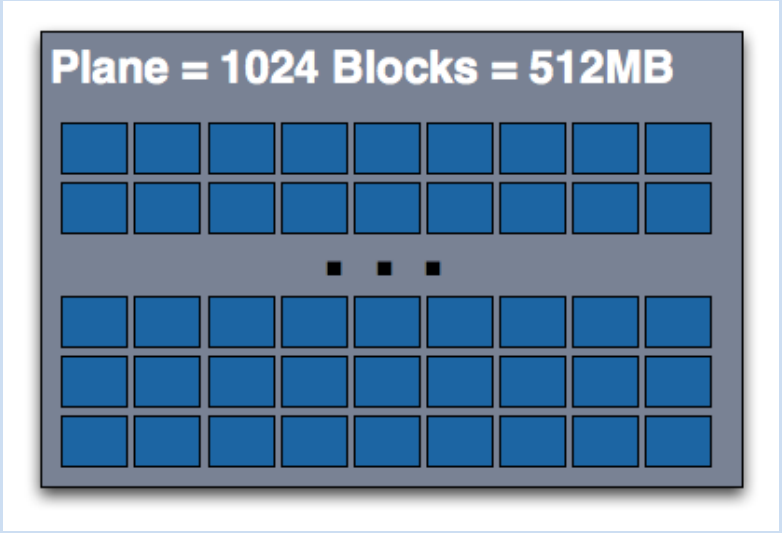
- **Blöcke: Zusammenfassung mehrerer Pages**
  - ein Block ist die kleinste Struktur einer SSD, die gelöscht werden kann
  - derzeit besteht ein Block zumeist aus 128 Pages á 4 KiB → 512 KiB Block
  - bei den neuen 25nm Flash Chips von Intel/Micron 256 Pages á 8 KiB → 2 MiB Block



Quelle: anandtech.com

1) Aufbau

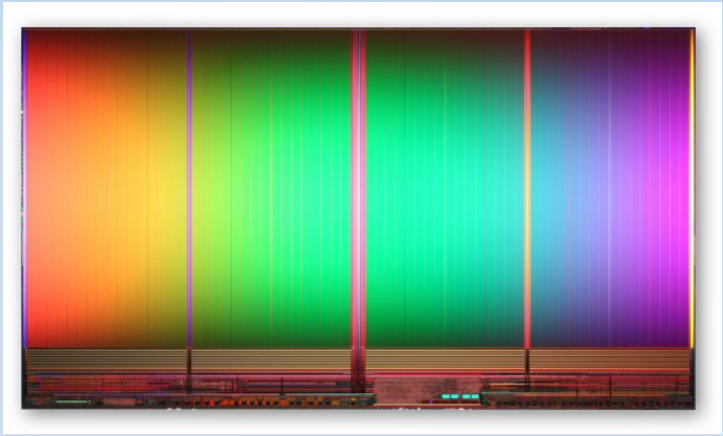
- **Planes**
  - mehrere Blöcke bilden eine Plane
  - zumeist 1.024 Blocks = 1 Plane
  - 25nm Intel/Micron:
    1 Plane = 2 GiByte



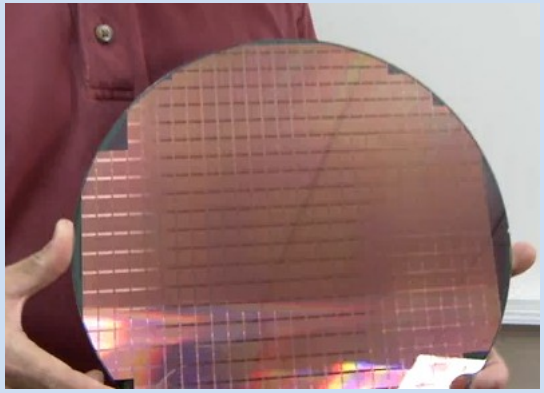
Quelle: anandtech.com

1) Aufbau

- **Dies**
  - mehrere Planes bilden ein Die
  - zumeist 4 Planes = 1 Die
  - 25nm Intel/Micron:
    1 Die = 8 GiByte
  - Beispiel eines Wafers
    mit zahlreichen Dies:



Quelle: http://www.intel.com/pressroom/archive/releases/20100201comp.htm



Quelle: Intel/Micron

1) Aufbau

- **TSOPs (thin small outline packages)**
  - mehrere Dies bilden ein TSOP

Quelle: Intel/Micron

  - 25nm Intel/Micron:
    8 Dies = 1 TSOP = 64 GiByte

1) Aufbau

- **SSDs**
  - mehrere TSOPs (zum Beispiel zehn) sind in einer SSD
  - ab Anfang 2011 vermutlich Kapazitäten bis zu 600 GByte

Quelle: maximumpc.com

Agenda

1) SSD Aufbau

2) **Schreibtechniken**

- **Spare Area**
- **Wear Leveling**
- **ATA TRIM**
- **Garbage Collection**
- **Secure Erase**
- **Lebensdauer (Endurance)**

3) Anwendungsbeispiele

4) Konfigurationstipps

2) Schreibtechniken

- **Spare Area**
  - typsicherweise zwischen 7% und 28% der Netto-Kapazität
  - z.B. 160 GByte sichtbar, tatsächliche Kapazität 160 GiByte (171,8 GByte → 11,8 GByte Spare Area)
  - Spare Area wird genützt für
    - Read/Modify/Write
    - Wear Leveling
    - Bad Block Replacement
- **Wear Leveling**
  - Flash Speicherzellen sind nur endlich oft beschreibbar
  - Wear Leveling verteilt die Abnutzung möglichst gleichmäßig auf alle Speicherzellen

2) Schreibtechniken

- **ATA TRIM**
  - Betriebssystem teilt der SSD mit welche Datenbereiche nicht mehr benötigt werden und gelöscht werden können
  - erhöht die Anzahl an gelöschten Blöcken und damit die Schreibperformance
  - muss unterstützt werden von
    - SSD
    - Betriebssystem
    - Dateisystem
- **Garbage Collection**
  - in Zeiten ohne I/O fasst der SSD-Controller teilweise beschriebene Blöcke zusammen
  - erhöht die Anzahl an gelöschten Blöcken

2) Schreibtechniken

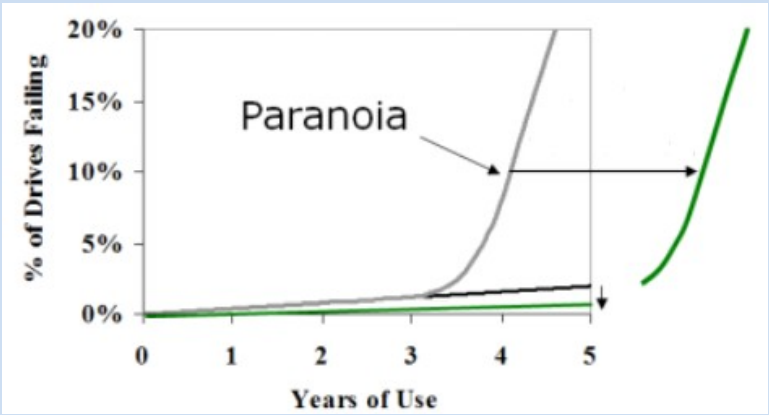
- **Secure Erase**
  - alle Daten gehen verloren
  - löscht alle Blöcke einer SSD
  - Löschspannung wird angelegt
  - danach wieder optimale Schreibperformance
  - empfehlenswert
    - wenn gebrauchte SSDs für neue Zwecke genutzt werden
    - nachdem Performance-Tests durchgeführt wurden

2) Schreibtechniken

- **Lebensdauer (Endurance)**
  - Bad Blocks (Ausfall von Blöcken)
    - je mehr p/e Zyklen stattfanden, umso länger dauert das Löschen eines Blockes
    - überschreitet die notwendige Zeit zum Löschen eine bestimmte Dauer, wird der Block als Bad Block markiert, und stattdessen ein Spare Block verwendet
    - kein Datenverlust, einzelne Bad Blocks kein Problem
  - Write Data Errors
    - RBER (raw bit error rate) – wird mittels ECC korrigiert
    - RBER steigt mit höherer Zahl an p/e Zyklen
    - ECC kann eine gewisse Anzahl an Fehlern korrigieren
    - UBER (uncorrectable bit error rate) muss niedrig gehalten werden (meist <1 Fehler pro 10^{15} bis 10^{16} Zugriffen)

2) Schreibtechniken

- **Lebensdauer (Endurance)**
  - Data Retention
    - Anzahl an Stunden (Tage/Jahre) wie lange Daten erhalten bleiben nachdem das Gerät außer Betrieb genommen wurde
    - ECC kann eine gewisse Anzahl an Fehlern korrigieren
    - retention time sinkt mit höherer Zahl an p/e Zyklen
  - Defects
- **alle NAND-Devices haben ein „Wearout Cliff“**
  - neue JDEC Standards (TBW – Terabytes written)



Quelle: Intel

Agenda

1) SSD Aufbau

2) Schreibtechniken

3) **Anwendungsbeispiele**

- **SSD als (kleines) Bootdevice**
- **SSD als Ersatz für Einzel-HDD**
- **SSDs im RAID-Verbund**
- **SSD als Cache**

4) Konfigurationstipps

3) Anwendungsbeispiele

- **SSD als Bootdevice**
  - geringe Zahl an p/e Zyklen,
    täglicher Turnover z.B. 0,1x
  - geringe SSD Kapazität reicht bereits (geringe Kosten)
  - kürzere Startzeiten
  - Programme starten subjektiv deutlich schneller
  - erhöht die Produktivität beim Arbeiten am PC

3) Anwendungsbeispiele

- **SSD als Ersatz für Einzel-HDD**
  - normale bis geringe Anzahl an p/e Zyklen, täglicher Turnover maximal 0,5x; meist deutlich weniger
  - mittlere/hohe SSD Kapazität erforderlich (mittlere bis hohe Kosten)
  - geringere Stromverbrauch und geringe Abwärme, da eine normale Festplatte damit entfällt
  - vor allem bei Notebooks stark im Kommen
    - erhöht Akku-Laufzeit
    - verringert Gewicht

3) Anwendungsbeispiele

- **SSDs im RAID-Verbund**
  - normale bis geringe Anzahl an p/e Zyklen, täglicher Turnover maximal 0,5x; meist deutlich weniger
  - ATA TRIM greift hier nicht, da SSD aufgrund von RAID-Initialisierung/Partiy ständig logisch voll ist

3) Anwendungsbeispiele

- **SSD als Cache**
  - große Zahl an p/e Zyklen,
    täglicher Turnover z.B. 10x
  - spezielles Augenmerk auf SSD Endurance
    - vergrößerte Spare-Area erhöht Lebensdauer
  - Beispiele
    - Adaptec MaxIQ
    - Cache für ZFS (Beispiel NexentaStor)

Agenda

1) SSD Aufbau

2) Schreibtechniken

3) Anwendungsbeispiele

4) **Konfigurationstipps**

- **AHCI aktivieren**
- **Secure Erase vor dem Produktiveinsatz**
- **ATA TRIM aktivieren**
- **Partitions- und Dateisystem-Alignment**
- **Over-Provisioning**

4) Konfigurationstipps

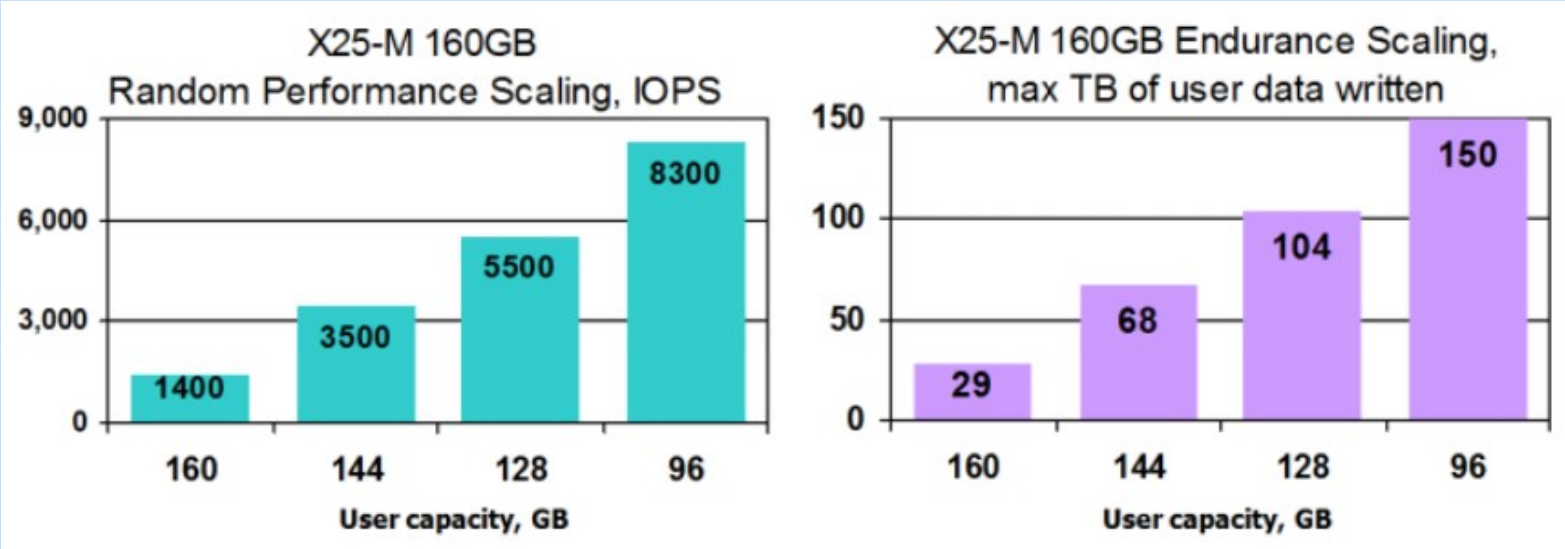
- **AHCI aktivieren**
  - NCQ (Native Command Queuing)
  - LPM (Link Power Management)
    - Device Initiated Interface Power Management (DIPM) aktivieren
- **Secure Erase vor dem Produktiveinsatz**
  - vor dem Partitionieren
  - löscht alle Blöcke der SSD
  - Blöcke aus Sicht des SSD Controllers „leer“
  - erhöht Performance

4) Konfigurationstipps

- **ATA TRIM aktivieren**
  - Linux ab 2.6.33 (z.B. Ubuntu 10.10)
- **Partitions- und Dateisystem-Alignment**
  - vor allem bei Generation-1 SSDs wichtig
  - Partitionierung unter Linux mit speziellen Parametern
    - fdisk -S 32 -H 32 /dev/sda
    - erste Partition erst bei Zylinder 2 starten lassen um richtiges Alignment zu gewährleisten
      (beim ersten Zylinder wird der erste Track – Head 0 – für den MBR freigehalten)

4) Konfigurationstipps

- **Over-Provisioning (vergrößern der Spare Area)**



Quelle: Intel

Fazit

Technologie der SSDs hat sich bereits sehr verbessert (intelligente Controller)

Durch geringe Strukturbreiten (25 nm) sinken die Preise/GByte

Lebensdauer/Zuverlässigkeit durch JEDEC Standard gut planbar

→ Bedeutung von SSDs und Flash-Speicher wie Fusion-io wird sich in den nächsten Jahren weiter deutlich steigern