I/O Performance optimieren

Werner Fischer, Technology Specialist Thomas-Krenn.AG

Thomas Krenn Herbstworkshop & Roadshow 2011

23.09. in Freyung
06.10. in Wien (A)
10.10. in Frankfurt
11.10. in Düsseldorf
12.10. in Hamburg
13.10. in Berlin
19.10. in München
20.10. in Zürich (CH)
25.10. in Prag (CZ)

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

Agenda

1) Speichermedien
2) Zugriffsmuster (Access Patterns)
3) Schnittstellen (Interfaces)
4) I/O Stack (Beispiel Linux)
5) Verbund mehrerer Speichermedien (RAID)
6) I/O Anforderungen
7) I/O Performance optimieren
8) Tips

Agenda

1) **Speichermedien**
2) Zugriffsmuster (Access Patterns)
3) Schnittstellen (Interfaces)
4) I/O Stack (Beispiel Linux)
5) Verbund mehrerer Speichermedien (RAID)
6) I/O Anforderungen
7) I/O Performance optimieren
8) Tips

1) Speichermedien

- **Festplatte**
  - gute sequential Performance
  - limitierte random Perfomance
  - random access time:

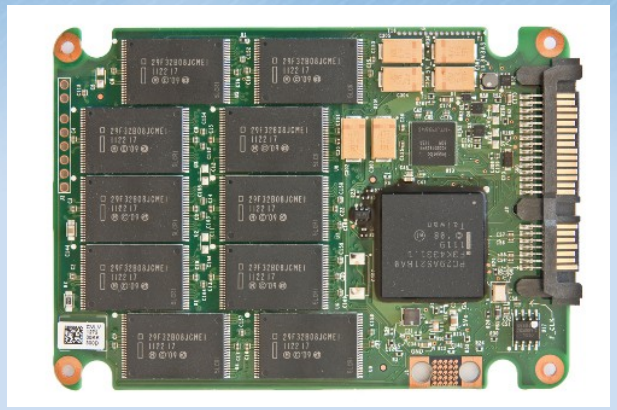
HDD	RPM	seek time*	latency	contr.overh.**	mittl.Zugriffsz.
WD 500 GB WD5002ABYS 3,5“	7.200	8,9 ms	4,2 ms	0,2 ms	13,3 ms
Seagate 400 GB ST3400755SS 3,5“	10.000	4,2 ms	3,0 ms	0,2 ms	7,4 ms
Fujitsu 300 GB MBA3300RC 3,5“	15.000	3,6 ms	2,0 ms	0,2 ms	5,8 ms

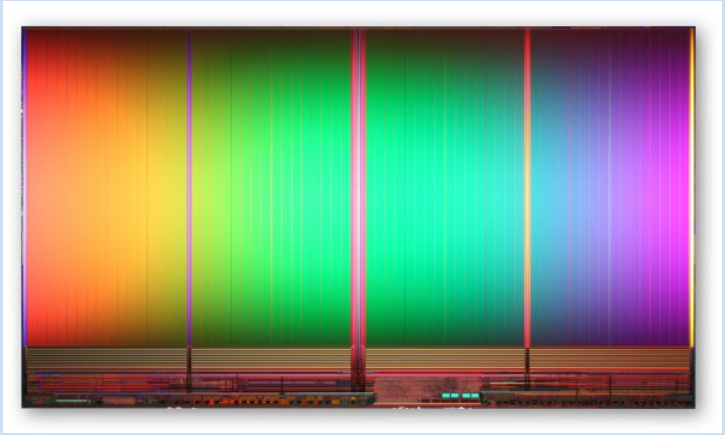
*) write seek time, Angabe laut Hersteller
**) Beispielannahme

Weitere Informationen: Präsentationen der Thomas Krenn Roadshow 2009

1) Speichermedien

- **Solid-State Drive (SSD)**
  - Aufbau:
    - Cells / Pages / Blocks
    - Planes / Dies / TSOPs / SSD
  - Spare Area
  - Wear Leveling
  - Write Amplification / Garbage Collection
  - ATA Trim



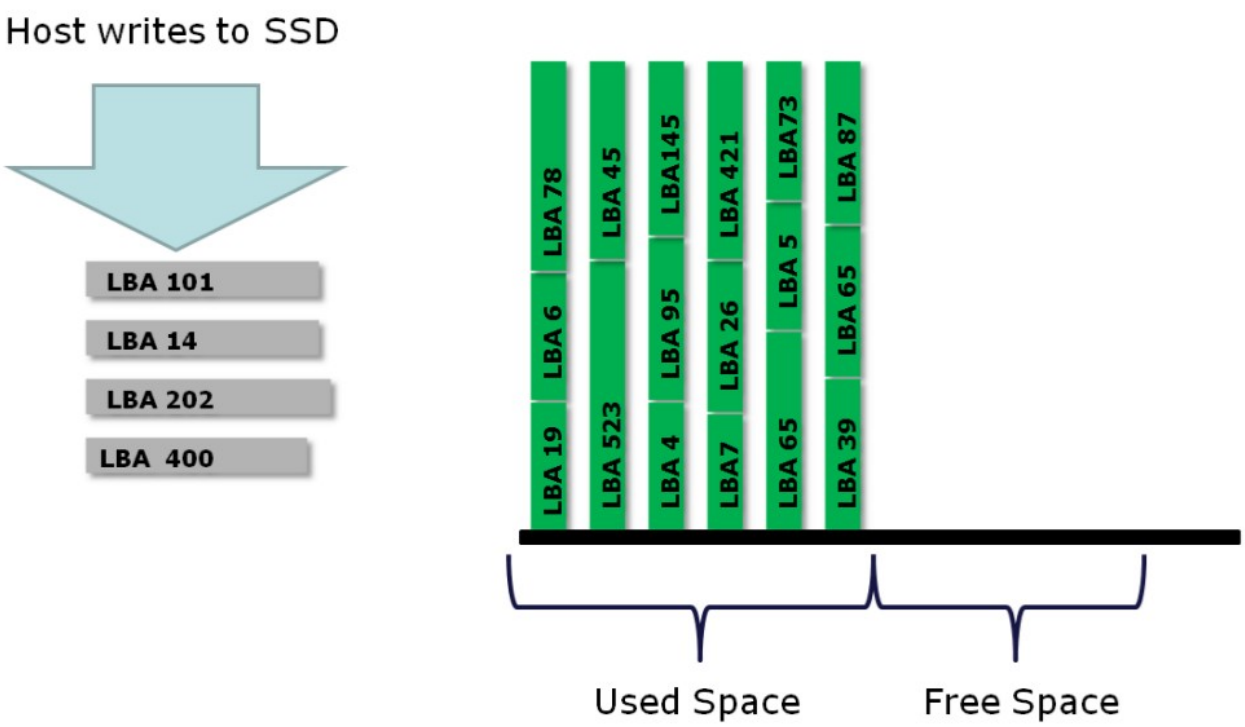


Quelle: http://www.intel.com/pressroom/archive/releases/20100201comp.htm

Weitere Informationen: Präsentationen der Thomas Krenn Roadshow 2010

1) Speichermedien

Drive Writes LBAs in Sequential Order
Even If They Are Random LBAs



9

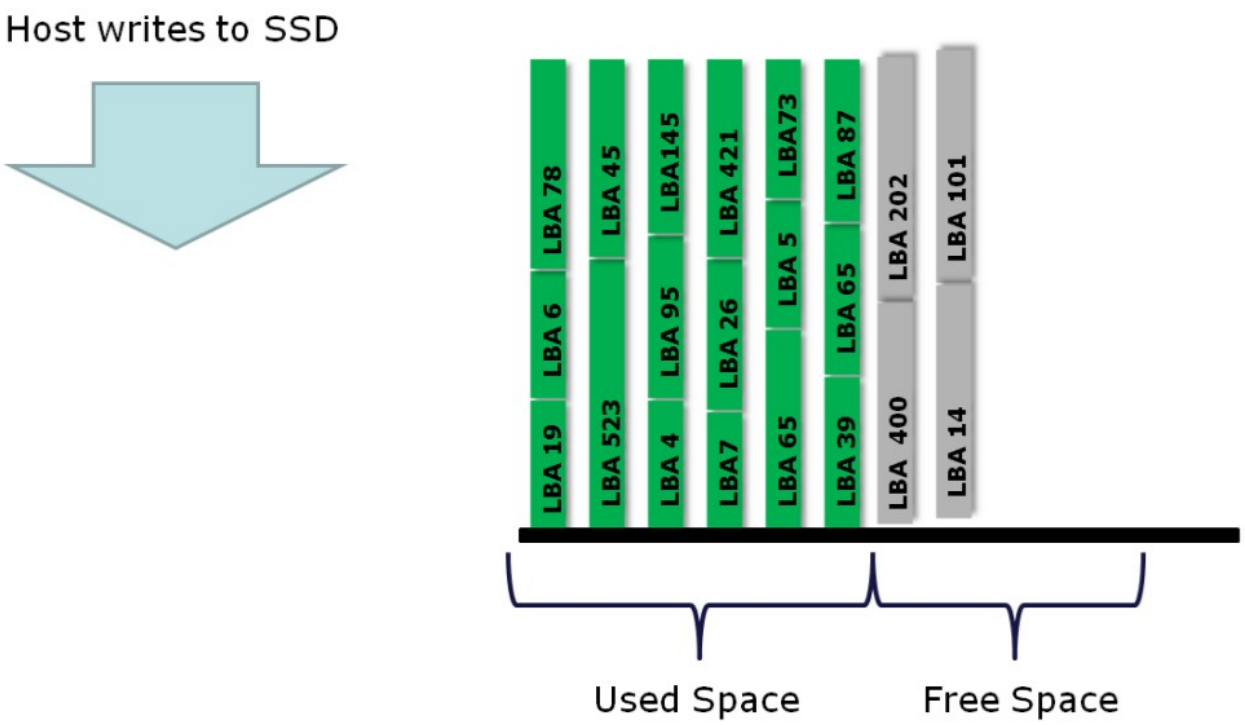
IDF2011
INTEL DEVELOPER FORUM

Quelle: Intel Developer Forum 2011

TK SERVER

1) Speichermedien

Drive Writes LBAs in Sequential Order
Even If They Are Random LBAs



IDF2011
INTEL DEVELOPER FORUM

9

Quelle: Intel Developer Forum 2011

TK SERVER

Agenda

1) Speichermedien
2) **Zugriffsmuster (Access Patterns)**
3) Schnittstellen (Interfaces)
4) I/O Stack (Beispiel Linux)
5) Verbund mehrerer Speichermedien (RAID)
6) I/O Anforderungen
7) I/O Performance optimieren
8) Tips

2) Zugriffsmuster (Access Patterns)

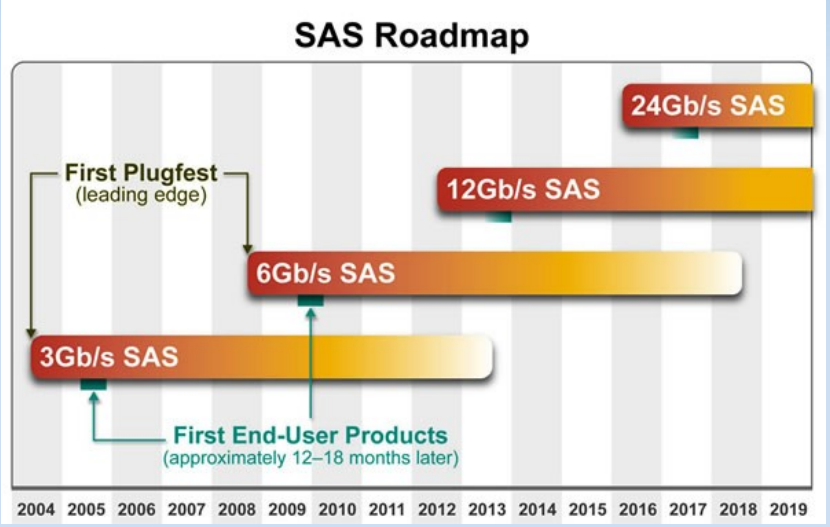
- **read / write**
  - read: Backup
  - mixed read/write: Datenbank Daten-Volume
  - write: Datenbank Transaction-Log, Restore
- **random / sequential**
  - random: Datenbank Daten-Volume
  - sequential: Datenbank Transaction-Log, Disk-Images, Backup
- **request size**
  - 4 KiB: Ext4/NTFS Blockgröße
  - 8 KiB: Exchange 2007
  - 256 KiB: Backup/Restore

Agenda

1) Speichermedien
2) Zugriffsmuster (Access Patterns)
3) **Schnittstellen (Interfaces)**
4) I/O Stack (Beispiel Linux)
5) Verbund mehrerer Speichermedien (RAID)
6) I/O Anforderungen
7) I/O Performance optimieren
8) Tips

3) Schnittstellen (Interfaces)

- **DAS (direct attached storage, block-based)**
  - SATA:
    - SATA 1,5 Gb/s
    - SATA 3 Gb/s
    - SATA 6 Gb/s
  - SAS:
    - 3 Gb/s SAS
    - 6 Gb/s SAS
  - PCIe, Zukunft: NVM EXPRESS



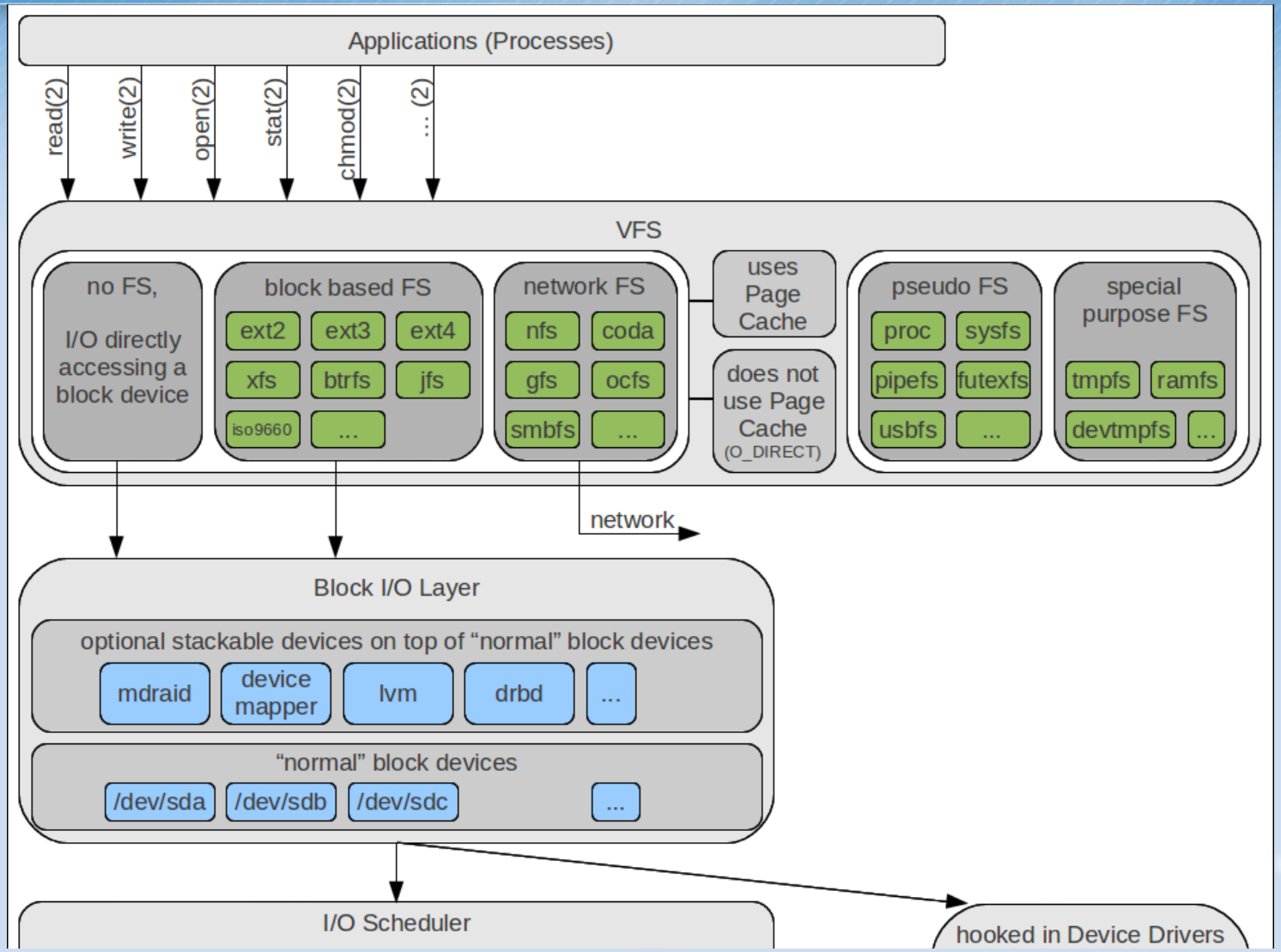
Quelle: http://www.scsita.org/sas_library/2011/06/serial-attached-scsi-master-roadmap.html

- **SAN (storage area network, block-based)**
  - iSCSI / FC
- **NAS (network attached storage, file-based)**
  - NFS / CIFS

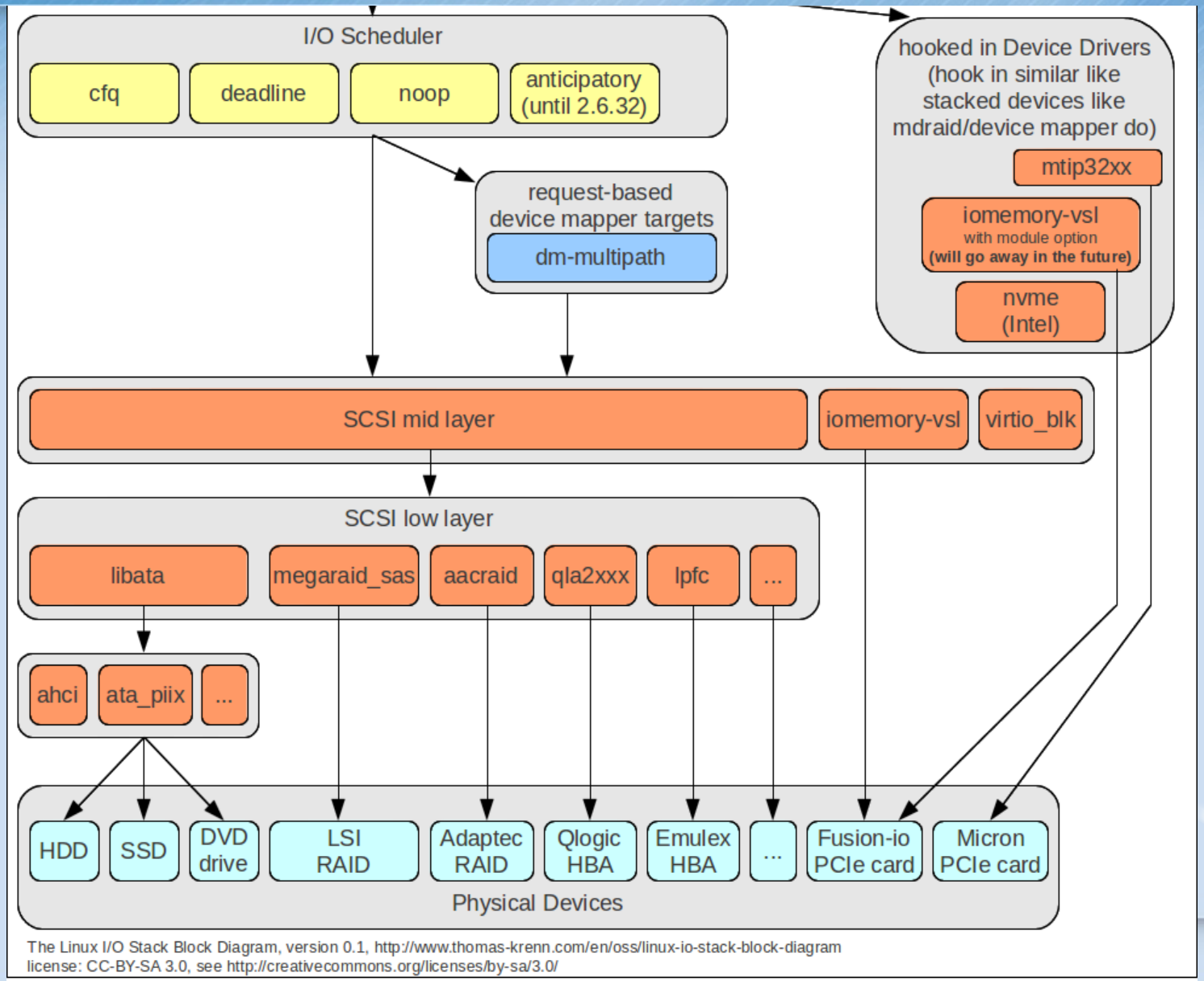
Agenda

1) Speichermedien
2) Zugriffsmuster (Access Patterns)
3) Schnittstellen (Interfaces)
4) **I/O Stack (Beispiel Linux)**
5) Verbund mehrerer Speichermedien (RAID)
6) I/O Anforderungen
7) I/O Performance optimieren
8) Tips

4) I/O Stack (Beispiel Linux)



4) I/O Stack (Beispiel Linux)

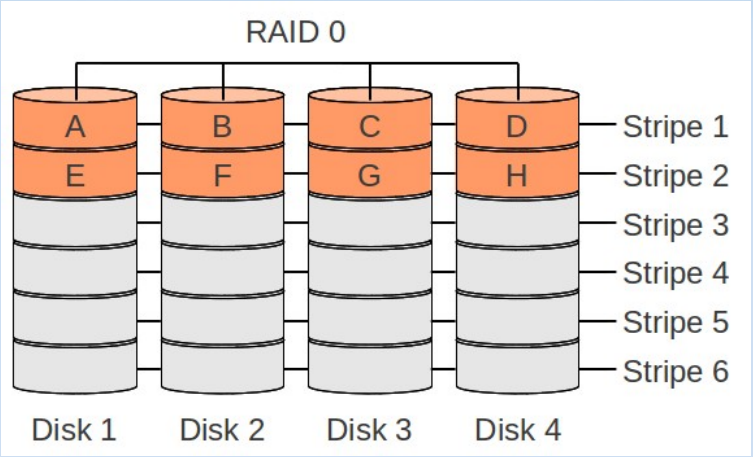


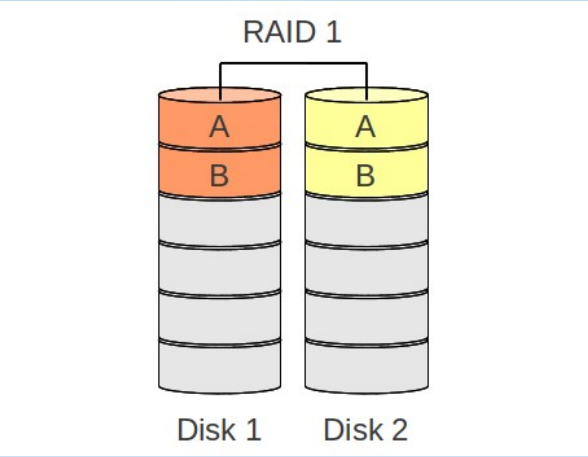
Agenda

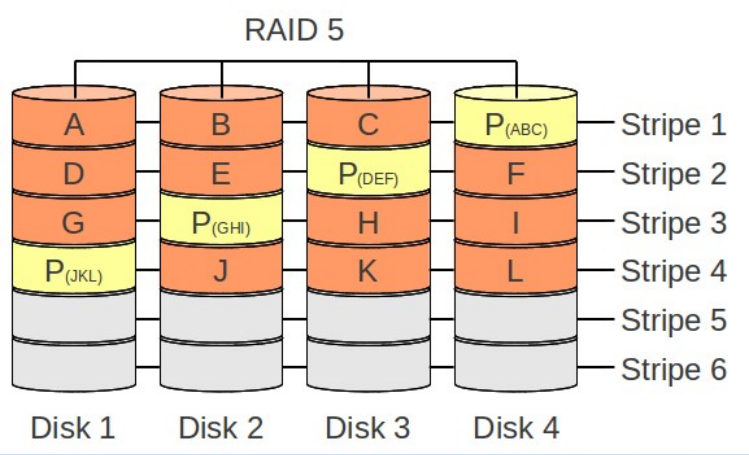
1) Speichermedien
2) Zugriffsmuster (Access Patterns)
3) Schnittstellen (Interfaces)
4) I/O Stack (Beispiel Linux)
5) **Verbund mehrerer Speichermedien (RAID)**
6) I/O Anforderungen
7) I/O Performance optimieren
8) Tips

5) Verbund mehrerer Speichermedien (RAID)

- **Beispiel: RAID 0 / RAID 1 / RAID 5**

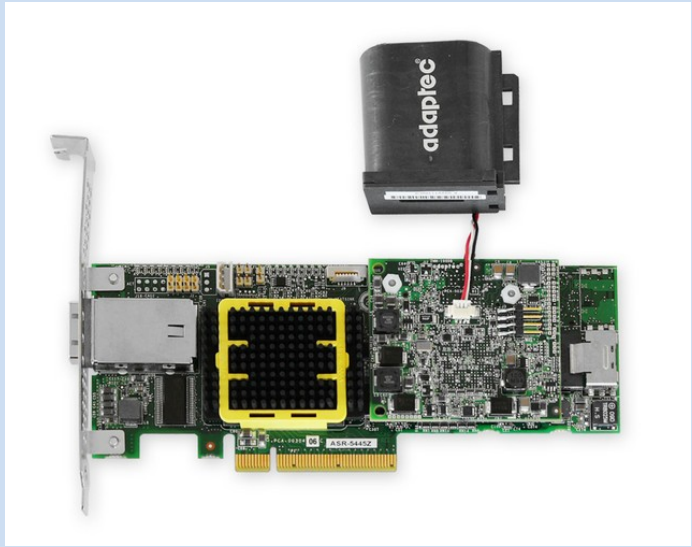






5) Verbund mehrerer Speichermedien (RAID)

- **RAID Typen**
  - Software RAID
  - Firmware/Driver RAID
  - Hardware RAID
    - Caches von HW-RAID Controller (gut für write / unnötig für read)
    - Cache-Protection
      - BBUs
      - Adaptec ZMCP
      - LSI CacheVault



Agenda

1) Speichermedien
2) Zugriffsmuster (Access Patterns)
3) Schnittstellen (Interfaces)
4) I/O Stack (Beispiel Linux)
5) Verbund mehrerer Speichermedien (RAID)
6) **I/O Anforderungen**
7) I/O Performance optimieren
8) Tips

6) I/O Anforderungen

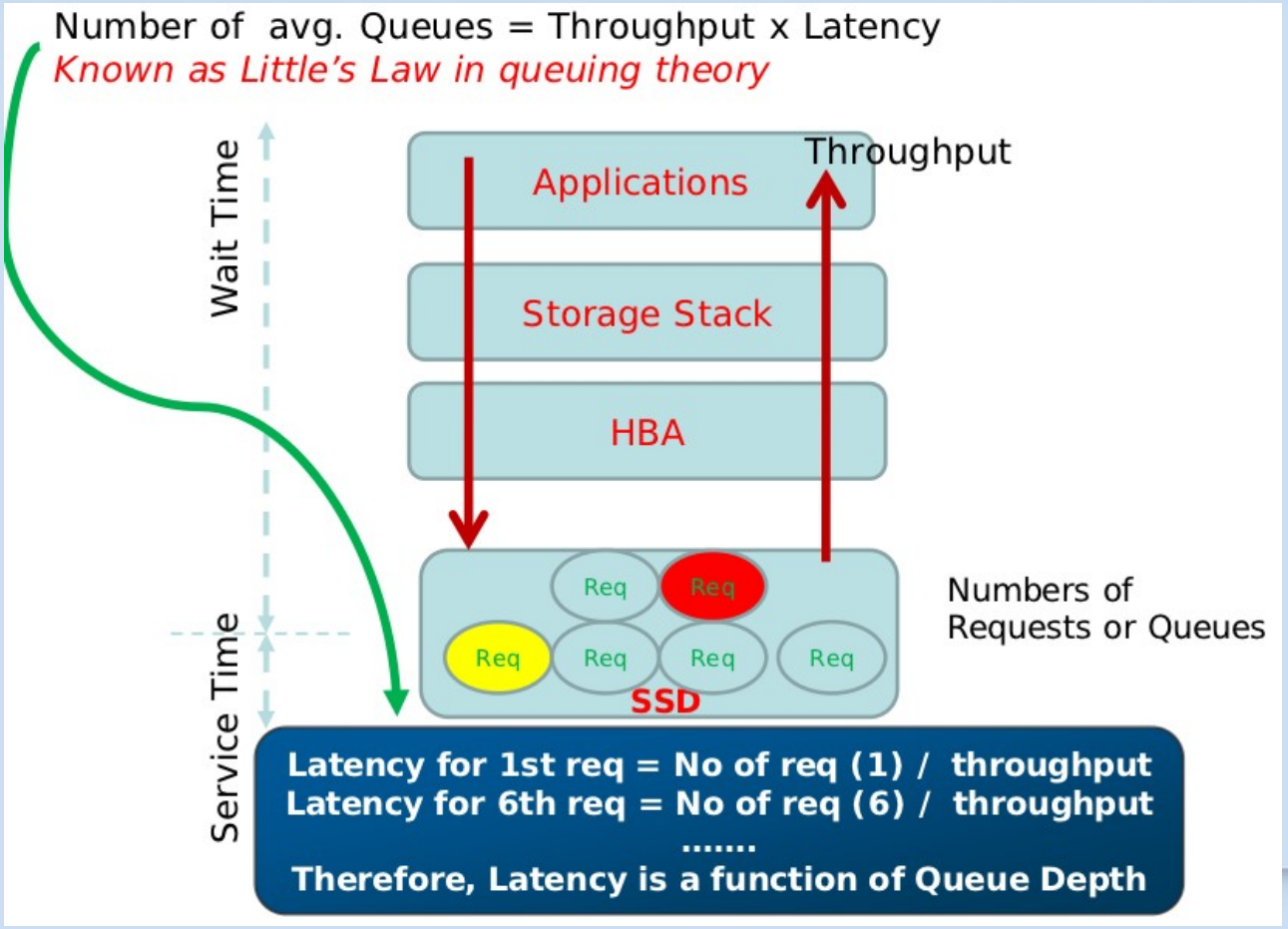
- **Transferrate / Datendurchsatz**
  - MB/s
  - Vergleich: Personen/h auf einer Strecke

- **Anzahl I/O Operationen pro Sekunde**
  - IOPS
  - Vergleich: Anzahl mögl. individueller Fahrten

- **dazu kommt: Latenz!**
  - Queue Depth
  - Vergleich: ab wie vielen Fahrzeugen fährt die Fähre los?

6) I/O Anforderungen

- **Latenz wird durch Länge der Warte-schlange (Queue) beeinflusst**



Agenda

1) Speichermedien
2) Zugriffsmuster (Access Patterns)
3) Schnittstellen (Interfaces)
4) I/O Stack (Beispiel Linux)
5) Verbund mehrerer Speichermedien (RAID)
6) I/O Anforderungen
7) **I/O Performance optimieren**
8) Tips

7) I/O Performance optimieren

- **verstehen**
  - Anforderungen
  - Aufbau Gesamtsystem
- **messen**
  - read/write
  - random/sequential
  - request size
- **optimieren**

Agenda

1) Speichermedien
2) Zugriffsmuster (Access Patterns)
3) Schnittstellen (Interfaces)
4) I/O Stack (Beispiel Linux)
5) Verbund mehrerer Speichermedien (RAID)
6) I/O Anforderungen
7) I/O Performance optimieren
8) **Tips**

TK SERVER

8) Tips

- **Alignment beachten, Bsp. SSD mit falschem Alignment:**

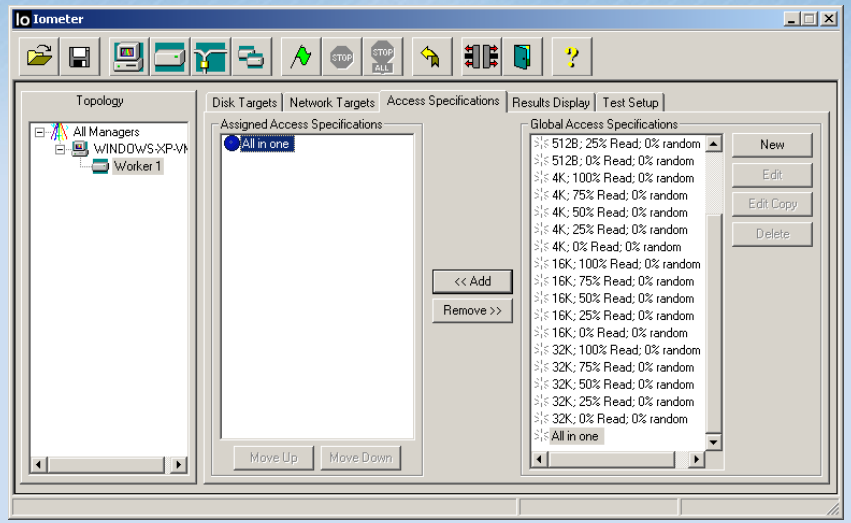
- **noatime / relatime**
- **SSDs:**
  - Over-Provisioning
  - Queue Depth – Abwägung IOPS/Latenz
  - SSD Sessions des Intel Developer Forum 2011: http://intel.com/go/idfsessions
- **Lastspitzen einplanen**

8) Tips (cont')

- **(Storage) Fehler einplanen, Auswirkungen durch**
  - RAID-Rebuild nach HDD-Ausfall
  - Ausfall eines Storage-Controllers
  - Ausfall eines Storage-Pfades bei Multi-Pathing
  - Leere Caches (Page Cache, DB Cache im RAM) nach Reboot
- **bei Neuplanung:**
  - Zugriffsmuster am Alt-System messen
  - Neues System anhand dieser Werte auslegen
  - Neues System vor Inbetriebnahme tunen (Zugriffsmuster mit IOmeter/fio simulieren)

8) Tips (cont')

- **Toolsammlung**
  - Lastgenerierung + Messung
    - IOmeter (Windows, Linux mit Windows-Client)
    - fio (Linux)
    - iozone (Linux)
  - Lastmessung
    - iostat
    - vmstat
    - Windows Performance Monitor



8) Tips (cont')

- **Thomas Krenn Wiki Artikel:**
  - RAID
  - RAID Controller Grundlagen
  - Linux Software RAID
  - Cache Einstellungen von RAID Controllern und Festplatten
  - Wartung der Battery Backup Unit (BBU/BBM) bei RAID-Controllern
  - Adaptec RAID Maintenance Best Practices
  - SSD Performance optimieren
  - SSD Over-Provisioning mit hdparm
  - ATA Trim
  - Ext4
  - ...

Optimale I/O Performance ist keine einmalige Konfiguration,

sondern ein kontinuierlicher Prozess.