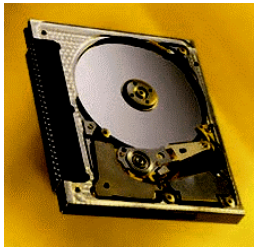
Disks: Agenda

- Festplatten
- IDE - Schnittstelle
- SCSI
- RAID
- Filecache/ OS-Strategien



Disks: IBM Microdrive

- 340 MB, kleiner als PCMCIA-II Karte, 16 Gramm

Disks: "iron oxide valley"

"I think Silicon Valley was misnamed. If you look back at the dollars shipped in products in the last decade there has been more revenue from magnetic disks than from silicon. They ought to rename the place Iron Oxide Valley"

Al Hoagland, One of the Pioneers of Magnetic Disks (1982)

[Hennessy & Patterson, Computer Architecture, 6.2]

Disks: Literatur

Friedhelm Schmidt: SCSI-Bus und IDE-Schnittstelle, Addison-Wesley 93

H.-P. Messmer PC-Hardwarebuch, Addison-Wesley 97

c't Plattenkarussell

c't SCSI-Einführung, Hefte 17/98/184, 18/98/144, 19/98/264

ATA-1 bis ATAPI-5 Spezifikationen

SCSI-1 bis SCSI-3 Spezifikationen

SCSI-3 MMC Spezifikation

www.seagate.com, www.quantum.com, www.storage.ibm.com

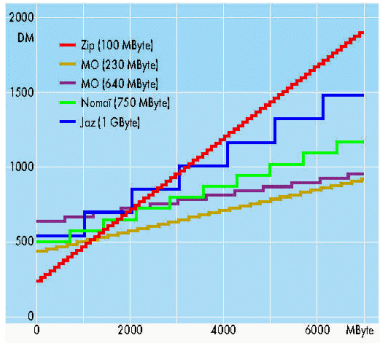
Disks: Plattenkarussell

Festplatten mit EIDE-Schnittstelle

	Kapazität	Drehzahl	Cache	Bauform	Random Access	Dauertransferrate Lesen min/mittel/max	Dauertransferrate Schreiben min/mittel/max	Gewichteter Mittelwert (Hdtune/h)	Interface	Lautgeräusch Ruhe	Lautgeräusch Betrieb
	[MByte]	[U/min]	[KByte]	[Zoll]	[ms]	[MByte/s]	[MByte/s]	[MByte/s]		[dBA/Sone]	[dBA/Sone]
								besser ▶			
Seagate											
ST31621A [33]	1548	3600	64 [4]	3,5/1	23,0/18,6	1,47/2,89/4,01	1,47/2,69/3,97	2,22	4	33,5/1,5	47,8/4,7
ST317242A Medalist17242	16447 [12]	5400	512	3,5/1	13,2/8,1	8,93/12,4/15,5	8,88/12,4/15,5	7,58	4, U4	36,1/1,7	42,0/3,1
ST32122A Medalist [5]	2014	4500	128 [4]	3,5/1	19,9/15,7	3,54/5,35/6,62	3,54/5,35/6,62	3,65	4, U2	37,7/2,0	47,7/4,7
ST32140A [28]	2015	5400	128	3,5/1	19,1/14,6	2,56/3,81/4,86	2,52/3,75/4,86	2,81	4	41,8/2,8	47,5/4,6
ST34321A Medalist4321 [17]	4103	5400	128	3,5/1	15,0/11,2	5,58/7,79/9,63	5,19/7,36/9,38	5,32	4, U2	33,8/1,3	38,9/2,4
ST34342A Medalist [10]	4103	4500	128	3,5/1	20,7/15,5	3,26/5,40/6,82	3,21/5,38/6,82	3,68	4, U2	39,1/2,7	49,0/5,6
ST36450A Medalist Pro [15]	6149	5400	448	3,5/1	16,7/11,1	5,08/6,90/8,65	4,69/6,72/8,65	4,68	4	38,4/2,1	45,5/4,1
ST36451A Medalist Pro 6451 [3]	6149	5400	448	3,5/1	16,5/11,2	5,04/6,85/8,59	5,04/6,85/8,59	4,91	4, U2	36,0/1,9	46,2/4,3
ST36530A Medalist Pro 6530 [11]	6208	7200	448	3,5/1	14,3/9,2	6,51/11,8/14,0	6,50/11,8/14,0	7,23	4, U2	35,8/1,9	47,9/4,3
ST36531A Medalist6531 [17]	6204	5400	128	3,5/1	14,7/10,6	5,00/7,77/9,62	4,93/7,70/9,62	5,43	4, U2	34,0/1,5	43,1/3,4
ST3660A [OEM HP] [1]	520	3800	120	3,5/1	24,9/24,7	1,11/1,79/2,22	1,17/1,79/2,22	1,31	3	40,7/2,8	48,3/5,2
ST38421A U4 8421	8056 [12]	5400	256	3,5/1	13,6/9,4	8,89/12,9/16,0	8,80/12,9/16,0	7,47	4, U4	30,5/1,1	42,3/3,4
ST38641A Medalist8641 [17]	8207 [12]	5400	128	3,5/1	15,3/10,7	5,07/7,83/9,59	4,96/7,72/9,59	5,40	4, U2	35,1/1,6	42,7/3,3
ST39140A Medalist Pro 9140 [11]	8693 [12]	7200	448	3,5/1	14,4/8,9	3,40/11,3/14,0	3,59/11,4/14,0	7,62	4, U2	36,8/2,1	50,7/5,1
ST51270A [29]	1223	5400	128	3,5/0,75	19,7/16,4	2,34/3,65/4,62	2,26/3,58/4,61	2,61	4	38,6/2,5	43,1/3,3
ST52520A Medalist Pro 2.5 [2]	2446	5400	112	3,5/0,75	16,0/12,0	4,45/6,74/8,56	4,63/6,74/8,56	4,64	4	35,6/1,8	46,1/3,9
Western Digital											
AC11200L Caviar [6]	1222	5200	256	3,5/1	18,6/15,3	4,79/7,06/9,11	4,10/7,03/9,13	4,72	4	34,1/1,3	40,5/2,5
AC21200 Caviar [39]	1222	5200	128	3,5/1	18,4/15,1	3,15/4,78/6,03	2,72/4,72/6,03	3,15	4	37,4/2,1	48,5/4,1
AC21600 Caviar [14]	1549	5200	128	3,5/1	18,3/14,8	4,01/5,68/7,22	2,85/4,79/6,79	3,39 [31]	4	36,8/1,9	49,4/4,0
AC22100 Caviar [31]	2014	5200	128	3,5/1	18,4/13,8	4,10/6,19/7,90	3,49/5,71/7,89	3,48	4	38,1/2,2	49,2/4,4
AC22500L Caviar [8]	2441	5200	256	3,5/1	18,5/13,6	4,85/7,15/9,23	4,62/7,08/9,24	4,69	4	38,8/1,6	48,9/3,6
AC23200L Caviar [9]	3098	5400	256	3,5/1	16,7/11,9	5,70/8,26/9,85	5,71/8,18/9,85	4,56	4, U2	36,6/1,4	48,7/4,0
AC24300L Caviar [7]	4112	5400	256	3,5/1	16,6/11,3	5,70/8,27/9,84	3,75/4,75/5,57	4,24	4, U2	37,2/2,0	42,9/3,2
AC26400B Caviar [9]	6149	5400	512	3,5/1	16,4/10,5	7,57/10,5/12,3	7,57/10,5/12,3	5,62	4, U2	32,5/1,1	45,7/3,4
AC28400R Caviar [22]	8064 [12]	5400	512	3,5/1	14,8/10,3	8,14/11,3/13,1	8,14/11,3/13,1	6,94	4, U4 [33]	31,5/1,1	47,2/3,7
AC29100D Expert [13]	8693 [12]	7200	1966	3,5/1	13,3/9,1	10,3/14,3/17,0	10,1/14,3/17,0	10,5	4, U4	41,0/2,5	44,1/3,5
AC31010B Caviar [9]	9671 [12]	5400	512	3,5/1	16,4/10,3	7,57/10,8/12,7	7,57/10,8/12,7	5,73	4, U2	33,8/1,3	48,9/4,1
AC313000R Caviar [22]	12417 [12]	5400	512	3,5/1	15,0/10,0	7,84/11,2/13,1	7,84/11,2/13,1	7,29	4, U4 [33]	33,6/1,5	48,2/3,5

Leerseite

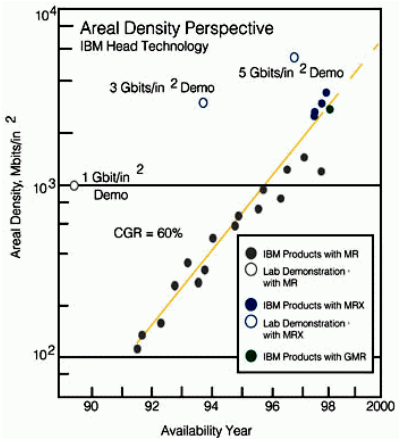
Disks: einige Wechselplatten



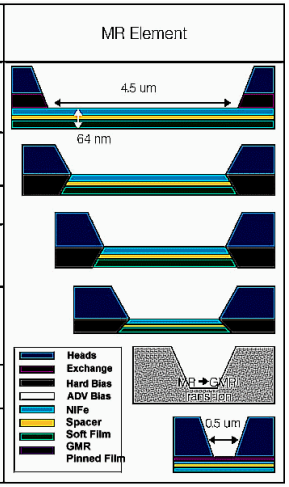
- diverse aktuelle Wechselplatten, magnetisch/magnetooptisch
- Kapazität vs. Performance vs. Kosten vs. Kosten/MB
- MO bietet extreme Datensicherheit, aber schlechtere Performance

Leerseite

Disks: Trend

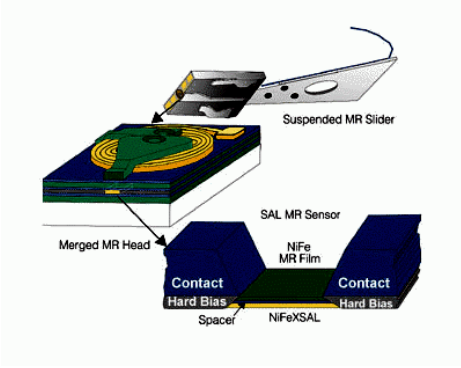


Year	Areal Density Gbits/in²
1991	0.132
1992	0.260
1993	0.354
1994	0.578
1995	0.829
	0.923
1996	1.32
1997	1.45
	2.64
1998	3.0-5.0
2000	10.0



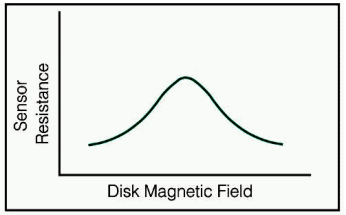
- 60% .. 70% / Jahr
  [www.storage.ibm.com]

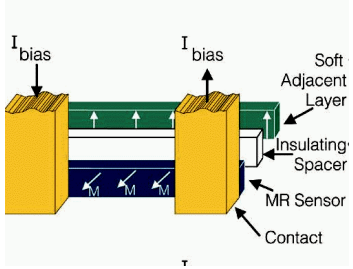
Disks: MR-Lesekopf: Aufbau

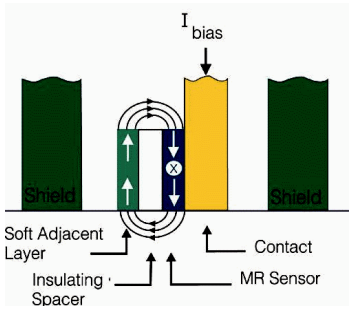


- Schreiben mit "normaler" Spule
- magnetoresistiver Lesekopf (MR)

Disks: MR-Lesekopf: Prinzip

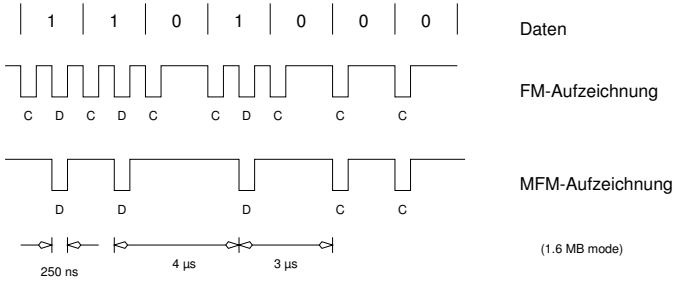






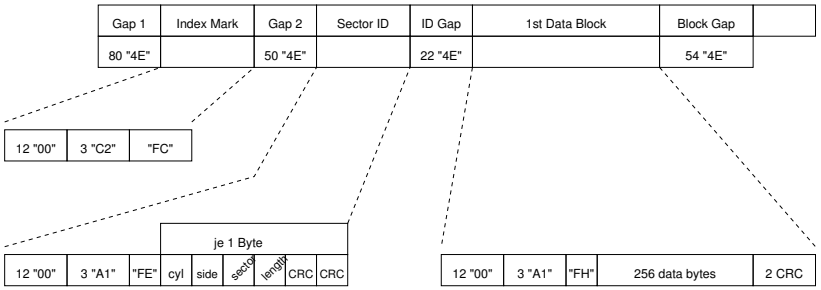
- [www.storage.ibm.com]

Disks: FM/MFM Aufzeichnung (Floppy)



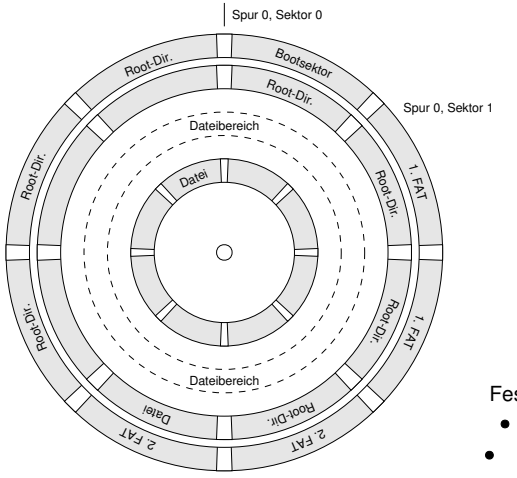
- Flusswechsel möglichst eng für hohe Speicherkapazität
- begrenzt durch Material, Lesekopf, oder Elektronik
- Frequenzmodulation verwendet Takt- und Datenimpulse
- MFM doppelte Kapazität
- Festplatten: Lauflängenkodierung (RLL) für höhere Kapazität

Disks: MFM Sektorformat



- keine separate Taktspur: selbsttaktend, spurführend
- muß Drehzahlschwankungen ausgleichen
- hohe Redundanz, spez. Taktmuster, CRC-Fehlerkorrektur
- Index-Markierungen für Spur/Sektornummer
- wird beim Formatieren erzeugt (nur Floppy)

Disks: Floppy-Sektorlayout



Floppy:

- Spur 0 außen:
- Bootsektor
- FATs
- Stammverzeichnis
- Dateien ab Spur 1

Festplatten:

- entsprechender Aufbau
- bad-sector management
- genaue Lage CHS unbekannt

Disks: Partitionssektor

Partitionssektor

Inhalt	Bytes	Adresse
Programm (Selbsttest, Aufruf des Bootsektors)	446	0000h
Partitionstabelle	64	01beh
Signatur aa55h	2	01feh / 01ff

Partitionstabelle

Inhalt		Offset
Partition 4	16	1beh
Partition 3	16	1ceh
Partition 2	16	1deh
Partition 1	16	1eeh

Partitionseintrag

Offset	Bytes	Inhalt
00h	1	Bootindikator 80=boot, 00=inaktiv
01h	3	Beginn der Partition
04h	1	System: 1=FAT12, 4=FAT16, 6=DOS32MB, ...
05h	3	Ende der Partition
08h	4	Startsektor rel. zum Anfang der Platte
0ch	4	Anzahl Sektoren der Partition

Kopf	z98 Sektor	Zylinder

(Beginn/Ende einer Partition)

Selbsttest ab Adresse 0000, verzweigt zum Bootsektor

Disks: Bootsektor

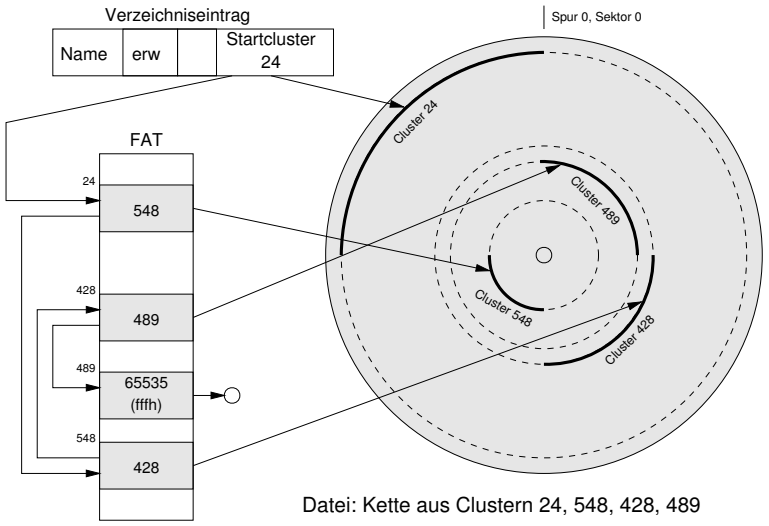
		Adresse
e9xxxxh oder ebxx90h	3 Bytes	00h / 0
OEM-Name und Nummer	8 Bytes	03h / 3
Bytes pro Sektor	2 Bytes	0bh / 11
Sektoren pro Cluster	1 Byte	0dh / 13
reservierte Sektoren (boot record)	2 Bytes	0eh / 14
Anzahl der FATs	1 Byte	10h / 16
Einträge im Root-Verzeichnis	2 Bytes	11h / 17
Anzahl der logischen Sektoren	2 Bytes	13h / 19
Medium-Desktriptor-Byte	1 Byte	15h / 21
Sektoren pro FAT	2 Bytes	16h / 22
Sektoren pro Spur	2 Bytes	18h / 24
Anzahl der Köpfe	2 Bytes	1ah / 26
Anzahl der verborgenen Sektoren	2 Bytes	1ch / 28
Programm zum Laden des Betriebssystems	Rest	1eh / 30

xxh : Start des Urladers

0f8h	Festplatte
0f0h	3.5", 72o KB
0f9h	3.5", 1.2 MB
0fah	3.5", 320 KB
0fbh	5.25", 640KB
. . .	

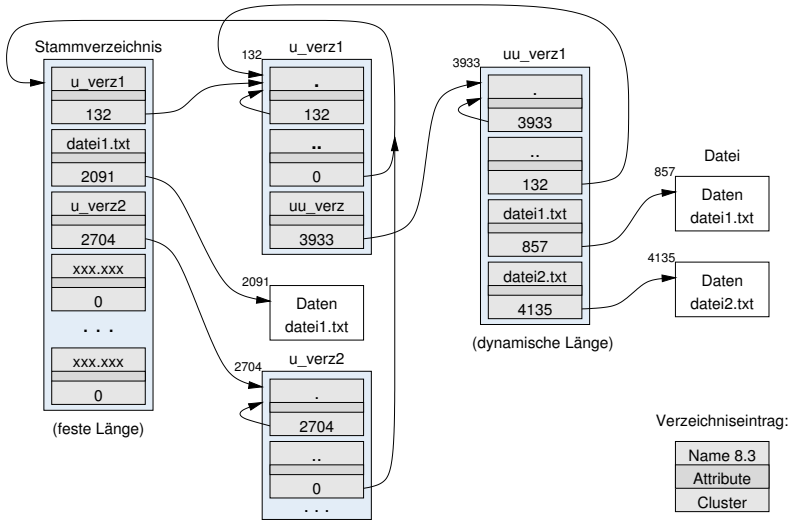
- erster Sektor der Partition (Kopf 0, Spur 0, Sektor 1)
- "Medium Descriptor Table" von 0bh .. 1eh
- ebxxxx: near jump xxxx / e9xx90: short jump xx nop

Disks: File Allocation Table (FAT)



Datei: Kette aus Clustern 24, 548, 428, 489

Disks: DOS-Verzeichnisstrutkur



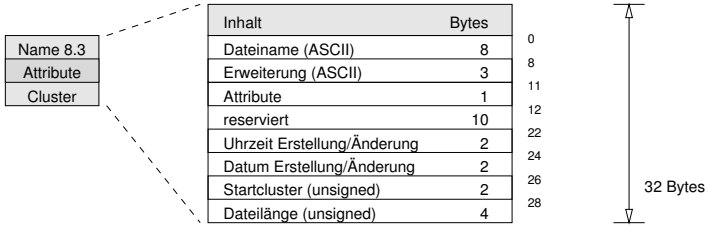
Disks: File Allocation Table

FAT-12	FAT-16	FAT-32	Bedeutung
000h	0000h	0000 0000h	frei
ff0h..ff6h	fff0h..fff6h	0fff fff0h..0fff fff6h	reserviert
ff7h	fff7h	0fff fff7h	defekter Sektor
ff8h..fffh	fff8h..ffffh	0fff fff8h..0fff ffffh	Ende der Clusterkette
xxxh	xxxxh	0xxx xxxxh	nächster Cluster der Datei
4077	65517	2^28 = 256M	max. Anzahl der Cluster

- geringe Anzahl der Cluster in FAT-16 führt zu riesigen Clustern:
- ungeeignet für große Platten / Vielzahl von kleinen Dateien

Kapazität	Clustergröße (FAT-16)
16..128 MB	2 KB (4 Sektoren)
128..256 MB	4 KB (8 Sektoren)
256..512 MB	8 KB (16 Sektoren)
512..1024 MB	16 KB (32 Sektoren)
1024..2048 MB	32 KB (64 Sektoren)

Disks: DOS-Verzeichniseintrag

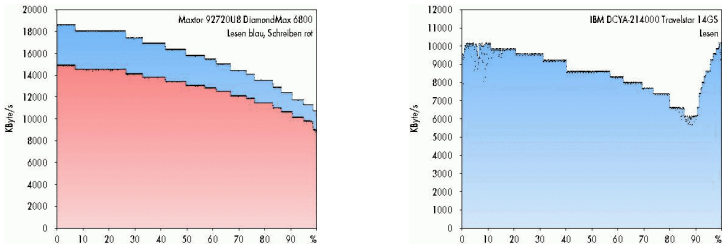


Dateiname:

- 8 Zeichen ASCII, 3 Zeichen Erweiterung
- Name ’2eh’ bzw. ’.’ bedeutet Verzeichnis, ".." das Stammverzeichnis
- Name ’e5h’ bedeutet "gelöscht"

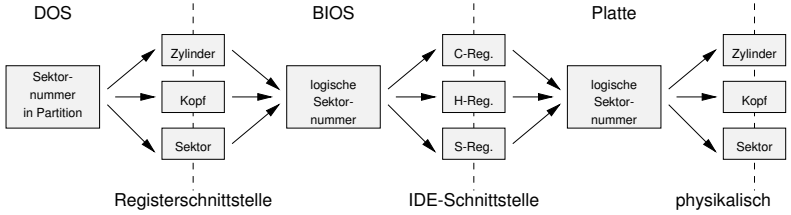
Disks: Zonenmessung

Anordnung der logischen Blöcke auf der Platte?



- c't-Messung: R/W-Transferraten als Funktion der Blockadresse
- viele Varianten möglich
- schnellste Zone (außen) meistens bei Adresse 0
- gibt es ein "optimales" Mapping?

Disks: BIOS/CHS/LBA-Adressierung



Adressierung von Daten auf einer Platte:

- CHS: Cylinder, Head, Sektor
- LBA, logical block addressing (fortlaufend ab 0)
- DOS/BIOS zu wenig Bits: Probleme bei 504M, 2G, 8G, ...

- herstellerspezifisches Mapping LBA - Sektor der Platte
- "Zonenmessung"

BIOS: 528 MByte Grenze (int13h)

3.2 The 528-megabyte barrier

BIOSs provide Int 13h services for accessing ATA drives from DOS. For conventional Int 13h the Cylinder-Head-Sector (CHS) values supplied to the Int 13h interface were passed to the drive without modification. This method of access allows "ill-behaved" applications to successfully access the drive, bypassing the BIOSs Int 13h interface. ATA drives support more than 1024 cylinders but the Int 13h interface is limited to 1024, this prevents the BIOS from accessing the full media by passing CHS values directly to the drive. Table 1 illustrates the limitations caused by the differences between the Int 13h and ATA maximum geometries.

Table 1 – Disk drive min/max

	BIOS	ATA	Limit
Max sectors/track	63	255	63
Max heads	256	16	16
Max cylinders	1024	65536	1024
Capacity	8.4 GB	136.9 GB	528 MB

This table illustrates how the conventional Int 13h interface with an 8.4 GB limit is restricted to 528 MB (63 * 16 * 1024 * 512). One solution to this problem is to address the drive using the Int 13h Extensions described in this technical report. Another solution is to create a false geometry that "fits" within Int 13h limitations, and also uses the full capacity of the drive. This capability is called geometric or drive translation. The translated geometry is applied in a manner that causes all sectors to maintain the same physical location on the media as when the drive is used in an untranslated environment. The Int 13h interface only has 10 bits for the cylinder, therefore Int 13h Fn 08h always returns the altered geometry information. This allows all DOS applications to function normally. Windows™ 3.11 and below functions normally when 32-bit disk access mode is disabled. A Windows™ driver which supports the geometry reported by Int 13h Fn 08h is required for 32-bit protected disk access mode.

BIOS: bit shifting

A simple bit-shift mapping scheme may create altered drive geometries. This method has the advantage of working with all ATA drives, including those drives which do not support LBA. A second advantage is that operation is fast and the code is small. The disadvantage of this method is that it lacks the flexibility to translate all geometries reported by a drive with a capacity less than 8.4 GB. However, drives which are ATA-2 (X3.279-1996) and above compatible will report geometries that may be translated. Annex D of ATA-2 or Annex B of ATA-3 and ATA/ATAPI-4 place limits on geometries for drives with less than an 8.4 GB capacity. The bit-shift method of translation manipulates the head and cylinder part of the geometry, but not the sectors per track. Table 2 describes the bit-shift translation capability:

Table 2 – Bit Shift Translation

Actual cylinders	Actual heads	Altered cylinder	Altered heads (see note)	Approx. size
1<C≤1024	1<H≤16	C=C	H=H	528 MB
1024<C≤2048	1<H≤16	C=C/2	H=H*2	1 GB
2048<C≤4096	1<H≤16	C=C/4	H=H*4	2.1 GB
4096<C≤8192	1<H≤16	C=C/8	H=H*8	4.2 GB
8192<C≤16384	1<H≤16	C=C/16	H=H*16	8.4 GB
16384<C≤32768	1<H≤8	C=C/32	H=H*32	8.4 GB
32768<C≤65536	1<H≤4	C=C/64	H=H*64	8.4 GB
NOTE – Value can not be greater than 255 in some Operating Systems.				

BIOS: LBA translation

Table 3 – LBA assist translation

Range	Sectors	Heads	Cylinders
1<X≤1,032,192	63	16	X/(1,008)
1,032,192<X≤2,064,384	63	32	X/(2,016)
2,064,384<X≤4,128,768	63	64	X/(4,032)
4,128,768<X≤8,257,536	63	128	X/(8,064)
8,257,536<X≤16,450,560	63	255	X/(16,065)
NOTE – X is the capacity of the drive, calculated by multiplying words 1, 3, and 6 of the IDENTIFY DEVICE data. This number may be different than the drive size reported by IDENTIFY DEVICE words 60 and 61.			

These two translation methods yield similar geometries in many cases. The difference between the two translations methods becomes apparent when a drive reports less than 63 sectors per track. The LBA assisted method always assigns a geometry with 63 sectors per track. The bit-shift method uses the sectors returned by

- evtl. andere Resultate als "bit-shifting"-Technik
- beide Varianten: bis max. 16 GByte
- beide Varianten: Platte nach BIOS-Wechsel evtl. nicht mehr lesbar

BIOS: extended read/write commands

4.2.2 Extended read

Entry:
AH - 42h
DL - Drive number
DS:SI - Disk address packet

Exit:
carry clear
AH - 0
carry set
AH - error code

This function transfer sectors from the device to memory. In the event of an error, the block count field of the disk address packet contains the number of good blocks read before the error occurred.

4.2.3 Extended write

Entry:
AH - 43h
AL - 0 or 1, write with verify off
2, write with verify on
DL - Drive number
DS:SI - Disk address packet

Exit:
carry clear
AH - 0
carry set
AH - error code

This function transfer sectors from memory to the device. If write with verify is not supported, this function rejects the request with AH=01h, CF=1. Function 48h is used to detect if write with verify is supported. In the event of an error, the block count field of the disk address packet contains the number of blocks written before the error occurred. AL also contains the values 0, 1, or 2. This function rejects all other values with AH=01h, CF=1

BIOS: extended BIOS translation

Table 8 – Device address packet

Offset	Type	Description
0	Byte	Packet size in bytes. Shall be 16 (10h) or greater. If the packet size is less than 16 the request is rejected with CF=1h and AH=01h. Packet sizes greater than 16 are not rejected, the additional bytes beyond 16 shall be ignored.
1	Byte	Reserved, must be 0
2	Byte	Number of blocks to transfer. This field has a maximum value of 127 (7Fh). A block count of 0 means no data is transferred. If a value greater than 127 is supplied the request is rejected with CF=1 and AH=01.
3	Byte	Reserved, must be 0
4	Double word	Address of transfer buffer. The is the buffer which Read/Write operations will use to transfer the data. This is a 32-bit address of the form Seg:Offset.
8	Quad word	Starting logical block address, on the target device, of the data to be transferred. This is a 64 bit unsigned linear address. If the device supports LBA addressing this value should be passed unmodified. If the device does not support LBA addressing the following formula holds true when the address is converted to a CHS value: $LBA = (C_1 * H_0 + H_1) * S_0 + S_1 - 1$ Where: C_1 = Selected Cylinder Number H_0 = Number of Heads (Maximum Head Number + 1) H_1 = Selected Head Number S_0 = Maximum Sector Number S_1 = Selected Sector Number For ATA compatible drives, with less than or equal to 15,482,880 logical sectors, the H_0 and S_0 values are supplied by WORDS 3 and 6 of the IDENTIFY DEVICE command.

- lineare 64-bit LBA-Adressierung

BIOS: extended BIOS detection

4.2.1 Check extensions present

Entry:
AH - 41h
BX - 55AAh
DL - Drive number

Exit:
carry clear
AH - Version of extensions
AL - Internal use only
BX - AA55h
CX - Interface support bit map (seeTable 9)
carry set
AH - error code (01h, Invalid Command)

Table 9 – Extension result buffer

Bit	Description
0	1 - Fixed disk access subset
1	1 - Drive locking and ejecting subset
2	1 - Enhanced disk drive support subset
3-15	Reserved, must be 0

This function is used to check for the presence of Int 13h extensions. If the carry flag is returned set, the extensions are not supported for the requested drive. If the carry flag is returned cleared, BX shall be checked for the value AA55h to confirm that the extensions are present. If BX is AA55h, the value of CX is checked to determine what subsets of this interface are supported for the requested drive. At least one subset must be supported. The version of the extensions is 21h. This indicates that the Int 13h extensions are compliant with this technical report.

- lineare 64-bit LBA-Adressierung

BIOS: extended BIOS device parameters

Table 4 – Standard device parameter table

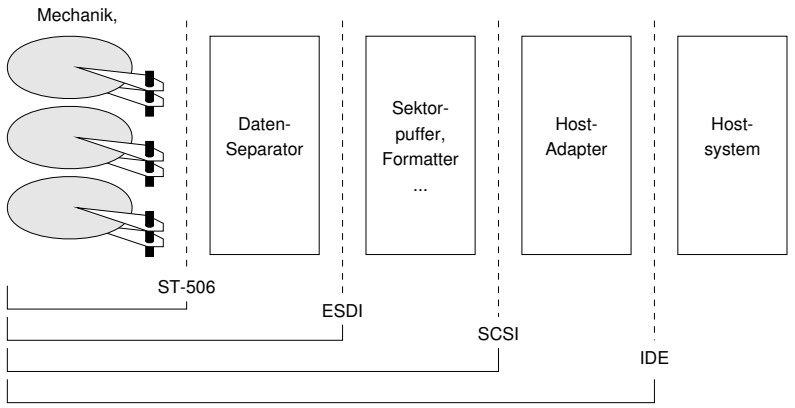
Byte	Type	Description
0-1	Word	Physical number of cylinders
2	Byte	Physical number of heads
3	Byte	Not Axh signature, indicates untranslated table
4	Byte	Reserved
5-6	Word	Precompensation (obsolete)
7	Byte	Reserved
8	Byte	Drive control byte
9-10	Word	Reserved
11	Byte	Reserved
12-13	Word	Landing zone (obsolete)
14	Byte	Sectors per track
15	Byte	Reserved

Table 5 – Translated device parameter table

Byte	Type	Description
0-1	Word	Logical cylinders, limit 1024
2	Byte	Logical heads, limit 256 (see note)
3	Byte	Axh signature, indicates translated table
4	Byte	Physical sectors per track, limit 63
5-6	Word	Precompensation (obsolete)
7	Byte	Reserved
8	Byte	Drive control byte
9-10	Word	Physical cylinders, limit 65536 (see note)
11	Byte	Physical heads , limit 16 (see note)
12-13	Word	Landing zone (obsolete)
14	Byte	Logical sectors per track, limit 63
15	Byte	Checksum, 2's complement of the 8 bit unsigned sum of bytes 0-14
NOTE – 0 indicates the maximum value. See table 2.		

- siehe "extended BIOS" specification

Disks: ST506 vs. SCSI vs. ATAPI



- IDE/ATA-Platte enthält kompletten Controller und Adapter

Disks: IDE-Schnittstelle

IDE	"integrated drive electronics"
EIDE	"enhanced IDE"
ATA	"AT attachment"
ATAPI	AT attachment packet interface

- Anschluss von Festplatten an den AT-Bus
- minimaler Hardwareaufwand des Interfaces (=billig)
- registerkompatible Variante eines WD ST506 Controllers
- vollständiger ST506-Controller in der Platte integriert (=IDE)

- mittlerweile standardisiert (ATA-1, 2, 3, 4, ATAPI, MMC, ...)
- Anschluss für Fest- und Wechselplatten, CD-Laufwerke, usw.
- derzeit fast immer im PC-Chipsatz integriert

- siehe ATAPI-5 Spezifikation

ATAPI: Signale

Table A.3 – 40-pin I/O connector interface signals

Signal name	Connector contact	Conductor		Connector contact	Signal name
RESET-	1	1	2	2	Ground
DD7	3	3	4	4	DD8
DD6	5	5	6	6	DD9
DD5	7	7	8	8	DD10
DD4	9	9	10	10	DD11
DD3	11	11	12	12	DD12
DD2	13	13	14	14	DD13
DD1	15	15	16	16	DD14
DD0	17	17	18	18	DD15
Ground	19	19	20	20	(keypin)
DMARQ	21	21	22	22	Ground
DIOW-:STOP	23	23	24	24	Ground
DIOR-:HDMARDY-:HSTROBE	25	25	26	26	Ground
IORDY:DDMARDY-:DSTROBE	27	27	28	28	CSEL
DMACK-	29	29	30	30	Ground
INTRQ	31	31	32	32	Obsolete (see note)
DA1	33	33	34	34	PDIAG-:CBLID-
DA0	35	35	36	36	DA2
CS0-	37	37	38	38	CS1-
DASP-	39	39	40	40	Ground
NOTE – Pin 32 was defined as IOCS16 in ATA-2, ANSI X3.279-1996.					

[ATAPI-5 Spec.]

- billiges 40-pol. Flachbandkabel
- Signale praktisch identisch mit den ISA-Bus Signalen
- seit kurzem auch 80-pol. Kabel für Ultra-DMA Modi

ATAPI: ATA-Register

Table F.4 – Register functions and selection addresses except PACKET and SERVICE commands

Addresses					Functions		
CS0-	CS1-	DA2	DA1	DA0	Read (DIOR-)	Write (DIOW-)	
N	N	x	x	x	Released	Not used	
					Control block registers		
N	A	N	x	x	Released	Not used	0
N	A	A	N	x	Released	Not used	4
N	A	A	A	N	Alternate Status	Device Control	6
N	A	A	A	A	Obsolete(see note)	Not used	7
					Command block registers		
A	N	N	N	N	Data	Data	0
A	N	N	N	A	Error	Features	1
A	N	N	A	N	Sector Count	Sector Count	2
A	N	N	A	A	Sector Number	Sector Number	3
A	N	A	N	N	Cylinder Low	Cylinder Low	4
A	N	A	N	A	Cylinder High	Cylinder High	5
A	N	A	A	N	Device/Head	Device/Head	6
A	N	A	A	A	Status	Command	7
A	A	x	x	x	Released	Not used	

Key:
A = signal asserted N = signal negated x = don't care
NOTE – This register is obsolete. It is recommended that a device not respond to a read of this address.

[ATAPI-5 Spec.]

- Host schreibt Parameter in Register 1-6
- Befehl starten durch Schreiben auf Register 7
- Datenübergabe nacheinander über das Data-Register 0

ATAPI: Befehle (Ausschnitt)

Command Name	Op Code	Type	Sub-clause
BLANK	A1h		6.1.1.
CLOSE TRACK/SESSION	5Bh		6.1.2.
FORMAT UNIT	04h		6.1.3.
INQUIRY	12h	M	SPC
LOAD/UNLOAD C/DVD	A6h	O	6.1.5.
MECHANISM STATUS	BDh	M	6.1.6.
MODE SELECT (6)	15h	M	SPC
MODE SENSE (10)	5Ah	M	SPC
MODE SENSE (6)	1Ah	M	SPC
PAUSE/RESUME	4Bh	A	6.1.7.
PLAY AUDIO (10)	45h	A	6.1.8.
PLAY AUDIO (12)	A5h	A	6.1.9.
PLAY AUDIO MSF	47h	A	6.1.10.
PLAY C/DVD	BCh	O	6.1.11.
PREVENT/ALLOW MEDIUM REMOVAL	1Eh	M	SPC
READ (10)	28h	M	SPC
READ BUFFER CAPACITY	5Ch		6.1.12.
READ C/DVD	BEh	O	6.1.13.
READ C/DVD MSF	B9h	O	6.1.14.
READ C/DVD RECORDED CAPACITY	25h	M	6.1.15.
READ DISC INFORMATION	51h		6.1.16.
READ DVD STRUCTURE	ADh		6.1.17.
READ HEADER	44h	M	6.1.18.
READ MASTER CUE	59h		6.1.19.
READ SUB-CHANNEL	42h	M	6.1.21.

ATAPI: Register für Packet-Command

Table F.5 – Register functions and selection addresses for PACKET and SERVICE commands

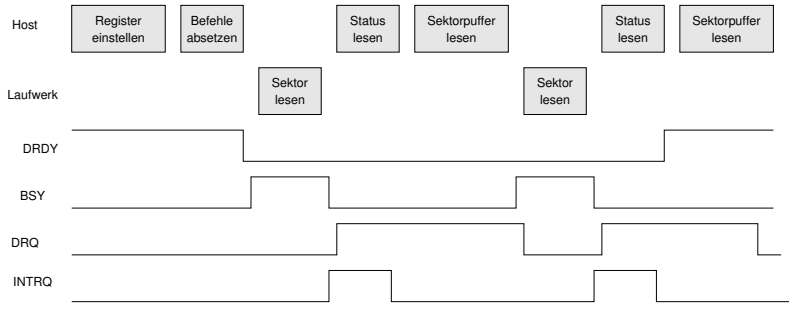
Addresses					Functions		
CS0-	CS1-	DA2	DA1	DA0	Read (DIOR-)	Write (DIOW-)	
N	N	x	x	x	Released	Not used	
					Control block registers		
N	A	N	x	x	Released	Not used	0
N	A	A	N	x	Released	Not used	4
N	A	A	A	N	Alternate Status	Device Control	6
N	A	A	A	A	Obsolete(see note)	Not used	7
					Command block registers		
A	N	N	N	N	Data	Data	0
A	N	N	N	A	Error	Features	1
A	N	N	A	N	Interrupt reason		2
A	N	N	A	A			3
A	N	A	N	N	Byte count low	Byte count low	4
A	N	A	N	A	Byte count high	Byte count high	5
A	N	A	A	N	Device select	Device select	6
A	N	A	A	A	Status	Command	7
A	A	x	x	x	Released	Not used	

Key:
A = signal asserted N = signal negated x = don't care
NOTE – This register is obsolete. A device should not respond to a read of this address.

[ATAPI-5 Spec.]

- CD/CDR/DVD haben andere Organisation als Festplatten
- Packet-Command definiert neue Bedeutung der Register
- Datentransfer wie bei normalen ATA-Befehlen

ATAPI: Prinzip PIO-Lesezugriff



- Laufwerk liest/schreibt jeweils ganzen Sektor
- PIO Host liest/schreibt jedes Datenwort einzeln
- DMA Datentransfer via DMA mit vollem Handshake
- Ultra-DMA DMA ohne Handshake, aber mit CRC

ATAPI: PIO-Modi 0 .. 4

Table 49 – PIO data transfer to/from device

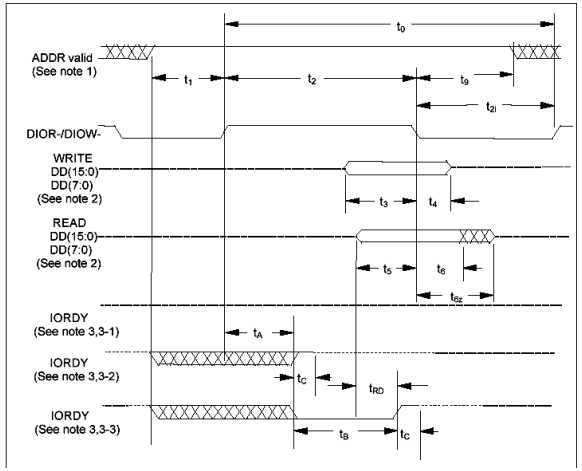
PIO timing parameters			Mode 0 ns	Mode 1 ns	Mode 2 ns	Mode 3 ns	Mode 4 ns	Note
t_0	Cycle time	(min)	600	383	240	180	120	1,4
t_1	Address valid to DIOR-/DIOW- setup	(min)	70	50	30	30	25	
t_2	DIOR-/DIOW-	(min)	165	125	100	80	70	1
t_{2i}	DIOR-/DIOW- recovery time	(min)	-	-	-	70	25	1
t_3	DIOW- data setup	(min)	60	45	30	30	20	
t_4	DIOW- data hold	(min)	30	20	15	10	10	
t_5	DIOR- data setup	(min)	50	35	20	20	20	
t_6	DIOR- data hold	(min)	5	5	5	5	5	
t_{6z}	DIOR- data tristate	(max)	30	30	30	30	30	2
t_9	DIOR-/DIOW- to address valid hold	(min)	20	15	10	10	10	
t_{RD}	Read Data Valid to IORDY active (if IORDY initially low after t_A)	(min)	0	0	0	0	0	
t_A	IORDY Setup time		35	35	35	35	35	3
t_B	IORDY Pulse Width	(max)	1250	1250	1250	1250	1250	
t_C	IORDY assertion to release	(max)	5	5	5	5	5	

NOTES –
1 t_0 is the minimum total cycle time, t_2 is the minimum DIOR-/DIOW- assertion time, and t_9 is the minimum DIOR-/DIOW- negation time. A host implementation shall lengthen t_2 and/or t_9 to ensure that t_0 is equal to or greater than the value reported in the devices IDENTIFY DEVICE data. A device implementation shall support any legal host implementation.
2 This parameter specifies the time from the negation edge of DIOR- to the time that the data bus is released by the device.
3 The delay from the activation of DIOR- or DIOW- until the state of IORDY is first sampled. If IORDY is inactive then the host shall wait until IORDY is active before the PIO cycle is completed. If the device is not driving IORDY negated at the t_A after the activation of DIOR- or DIOW-, then t_5 shall be met and t_{RD} is not applicable. If the device is driving IORDY negated at the time t_A after the activation of DIOR- or DIOW-, then t_{RD} shall be met and t_5 is not applicable.
4 Mode may be selected at the highest mode for the device if CS(1:0) and AD(2:0) do not change between read or write cycles or selected at the highest mode supported by the slowest device if CS(1:0) or AD(2:0) do change between read or write cycles.

[ATAPI-5 Spec.]

- Protokoll/Handshake immer gleich, unterschiedliche Wartezeiten

ATAPI: PIO Waveforms



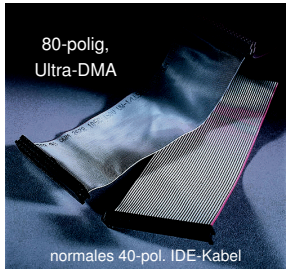
[ATAPI-5 Spec.]

- Host kontrolliert und initiiert alle Transfers

ATAPI: Ultra-DMA

- aktuelles, derzeit schnellstes Übertragungsverfahren
- Ultra-DMA/66 bis 66 MB/s

- Sender (Host/Platte) schickt Daten und Strobe-Impulse
- reduziertes Handshake
- dafür CRC-Fehlerkorrektur
- erfordert neues 80-pol. Kabel
- Anordnung abwechselnd Daten/Masseleitung



ATAPI: Ultra-DMA Waveforms

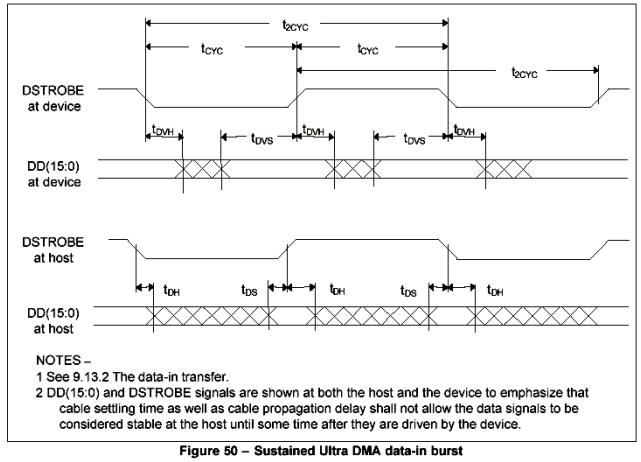


Figure 50 – Sustained Ultra DMA data-in burst

[ATAPI-5 Spec.]

- kein Handshake, jeweiliger Sender steuert Daten und Strobe

ATA: Marktbedeutung

126 Million Units and 87%
ATA must be doing something right!

Mobile+Desktop represent 126 MU in '98 and 87% of shipments. Category dominated by ATA.

Projections do not forecast any substantial change in the mix

Disk Drive Unit Shipments (in thousands)**

	Shipments	Forecast			
	1998	1999	2000	2001	2002
Mobile Drives	17846	20990	24340	28215	32600
Desktop Drives	108628	125646	143780	163180	184200
Server Drives	18493	21718	25700	30550	36130
Total	144967	168354	193820	221945	252930

Intel Developer Forum Spring 2000

****1999 Disk/Trend report at IIST Lk. Arrowhead**

Intel Labs

ATA: Serial-ATA

parallele Datenübertragung problematisch:

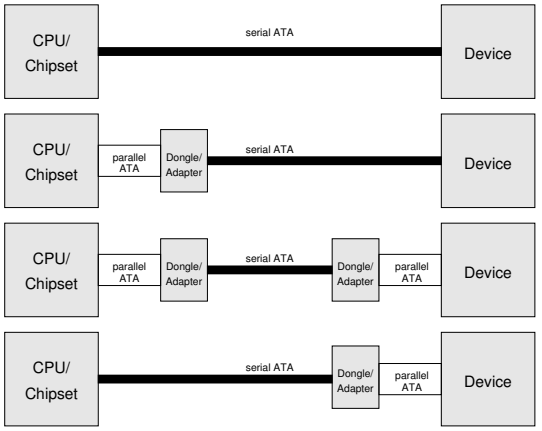
- teure Kabel
- Skew-Probleme
- höhere Taktraten als 100 MHz schwierig

=> Umstellung auf serielle Datenübertragung "Serial-ATA"

- Beibehalten des ATAPI-Befehlssatzes
- volle Kompatibilität
- Unterstützung durch alle großen Hersteller

- bei Bedarf "Dongles" zur Parallel/Seriell-Umwandlung

ATA: Serial-ATA Dongles



- bei Bedarf "Dongles" zur Parallel/Seriell-Umwandlung
- alte Hardware kann weiter genutzt werden, einfache Migration

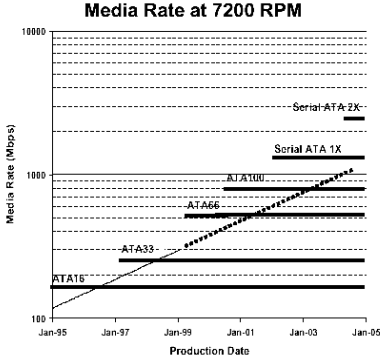
ATA: Serial-ATA Roadmap

Another Driving Factor

Interface rate driven to change with media rate

- **Minimizes buffering problems**

An ATA133 transition seems unnecessary



Intel Developer Forum Spring 2000

Interface takes a couple years to develop & deploy so some degree of developing in anticipation of the need is prudent

Intel Labs

SCSI: Übersicht

SCSI := Small Computer Systems Interface

- hervorgegangen aus "Shugart Associates SI"
- standardisiert als SCSI-I, SCSI-II, SCSI-III
- Einsatz in PCs (Server), Mac, Workstations

- keine reine Festplattenschnittstelle
- sondern universeller Bus für Peripheriegräte ("Targets")
- z.B. Bandlaufwerke, Scanner, Musiksynthesizer, ...
- 8-bit parallel (wide-SCSI mit 16-/32-bit)

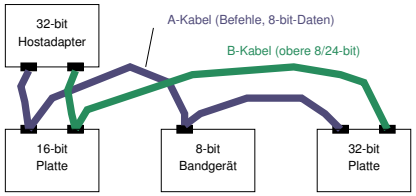
- "Hostadapter" steuert den Bus
- komplexe Befehle und Arbitrierung
- flexibler, aber auch teuer und komplexer als EIDE/ATAPI

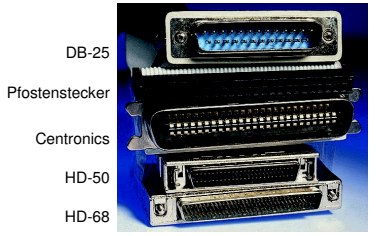
- Praxistips in der Artikelserie in ct 17-19/98

SCSI: Varianten

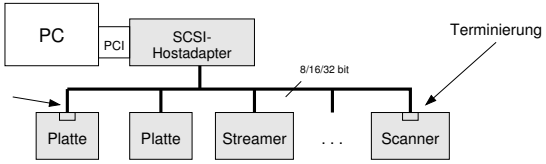
- Befehlssätze: SCSI-1, SCSI-2, SCSI-3
- Busbreite: normal 8-bit, wide-SCSI 16-bit und 32-bit
- Bustiming: SCSI-1 bis 5 MB/s, Fast 10 MB/s, Ultra 20 MB/s
- alle Kombinationen, z.B. U2W = Ulta-Wide SCSI-2

- alle Gerätevarianten miteinander kombinierbar
- insbesondere auch normale und wide-SCSI Geräte





SCSI: Grundlagen



- Bus mit 8 Geräten (LUN 0..7), inklusive Controller
- Gerätenummer per Schalter eingestellt (nicht automatisch!)
- komplexe Regeln zur Verkabelung (Terminierung, Abstände)

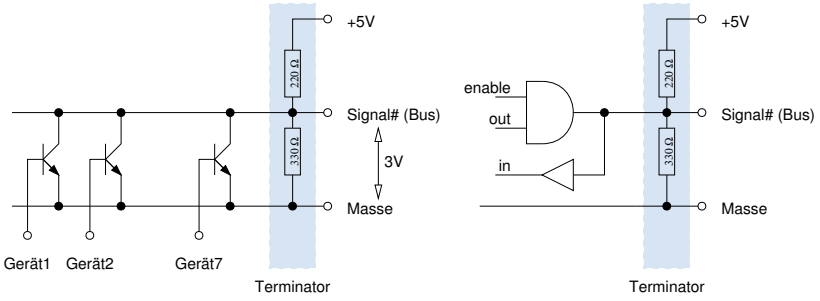
- parallele Datenübertragung, 8-bit oder (wide) 16/32-bit
- aufwendiges Busprotokoll mit Arbitrierung und split-transactions
- Geräte handeln die jeweils bestmögliche Geschwindigkeit aus
- langsame Geräte stören schnelle Geräte nicht

SCSI: Signale

Signal name	Connector contact number Set 2	Connector contact number Set 1	Cable conductor number	Connector contact number Set 1	Connector contact number Set 2	Signal name
GROUND	1	1	1	2	26	-DB(0)
GROUND	2	3	3	4	27	-DB(1)
GROUND	3	5	5	6	28	-DB(2)
GROUND	4	7	7	8	29	-DB(3)
GROUND	5	9	9	10	30	-DB(4)
GROUND	6	11	11	12	31	-DB(5)
GROUND	7	13	13	14	32	-DB(6)
GROUND	8	15	15	16	33	-DB(7)
GROUND	9	17	17	18	34	-DB(P)
GROUND	10	19	19	20	35	GROUND
GROUND	11	21	21	22	36	GROUND
RESERVED	12	23	23	24	37	RESERVED
OPEN	13	25	25	26	38	TERMPWR
RESERVED	14	27	27	28	39	RESERVED
GROUND	15	29	29	30	40	GROUND
GROUND	16	31	31	32	41	-ATN
GROUND	17	33	33	34	42	GROUND
GROUND	18	35	35	36	43	-BSY
GROUND	19	37	37	38	44	-ACK
GROUND	20	39	39	40	45	-RST
GROUND	21	41	41	42	46	-MSG
GROUND	22	43	43	44	47	-SEL
GROUND	23	45	45	46	48	-C/D
GROUND	24	47	47	48	49	-REQ
GROUND	25	49	49	50	50	-I/O

- 8-bit SCSI, entsprechend mehr Datenleitungen für Wide-SCSI

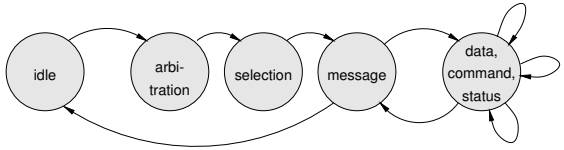
SCSI: Signale und Terminierung



- 8-bit SCSI hat 18 Signale auf 50-poligem Kabel
- Signale active-low mit open-Collector Schaltung: kurzschlußfest (!)
- ausgeschaltete Geräte stören den Bus nicht (!)
- Terminator zieht die Leitung auf "inaktiven" high-Pegel
- Terminierung nur an den beiden Endes des Busses

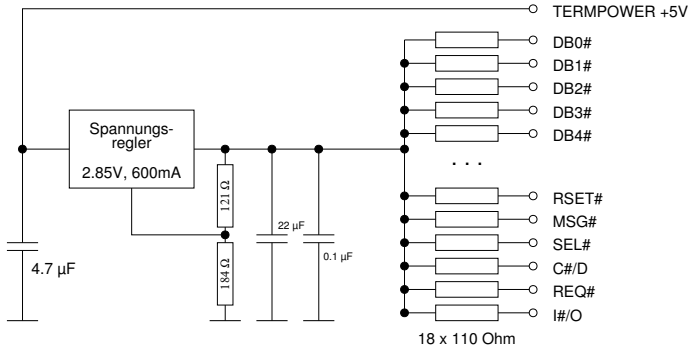
SCSI: Protokoll

- kompliziertes Mehrphasen-Busprotokoll:



- jeder Datentransfer erfordert die Arbitration-Phasen
- Datenphase erlaubt effiziente Burst-Transfers
- trotzdem beträchtlicher Overhead (H&P: 1 ms pro Transfer)
- mit gleicher Platte langsamer als ATAPI (aber flexibler)
- Arbitrierung/Befehlsübertragung immer 8-bit, asynchron
- Details siehe SCSI Spezifikation

SCSI: aktive Terminierung



- höhere Übertragungsrate erfordert Unterdrückung von Reflektionen
- geforderte Leitungsimpedanz 100..132 Ohm
- mit Spannungsregler / Konstantstromquelle

SCSI: SCSI-Befehlssatz

Table N.3 - Commands Common to all SCSI Devices

Command Name	Operation Code	SCSI-3 Type	SCSI-3 Ref Std
CHANGE DEFINITION	40h	O	
COMPARE	39h	O	
COPY	18h	O	
COPY AND VERIFY	3Ah	O	
INQUIRY	12h	M	
LOCK/UNLOCK CACHE	36h	O	
LOG SELECT	4Ch	O	
LOG SENSE	4Dh	O	
MODE SELECT (10)	55h	O	
MODE SELECT (6)	15h	M	
MODE SENSE (10)	5Ah	M	
MODE SENSE (6)	1Ah	M	
PREFETCH	34h	O	
PREVENT/ALLOW MEDIUM REMOVAL	1Eh	M	
READ (10)	28h	M	
READ (12)	A8h	O	
READ (6)	08h	O	
READ BUFFER	3Ch	O	
READ LONG	3Eh	O	
RECEIVE DIAGNOSTIC RESULTS	1Ch	O	
RELEASE (10)	57h	M	
RELEASE(6)	17h	O	
REQUEST SENSE	03h	M	
RESERVE(10)	56h	M	
RESERVE(6)	16h	O	
SEEK (10)	2Bh	M	
SEEK (6)	0Bh	M	
SEND DIAGNOSTIC	1Dh	M	
SET LIMITS (10)	33h	O	
SET LIMITS (12)	B3h	O	
START/STOP UNIT	1Bh	M	

Key: M = command implementation is mandatory
O = command implementation is optional

- für alle SCSI-Geräte
- zusätzliche Erweiterungen für Platten, Scanner, ...
- Standard: SCSI-3 MMC "multi media commands"

[SCSI-3 MMC spec]

SCSI: SCSI-3 MMC

"MultiMedia Command Set"

- standardisierte Befehlssatzerweiterung für SCSI
- insbesondere für CD/CDR/DVD/DVDR-Geräte:
  - digitales Auslesen von Audio-Tracks ("grabbing")
  - Ansteuerung von digitalen Audio-Ausgängen
  - Ansteuerung / Kalibrierung von CDR/DVD-Brennern
  - Unterstützung für das CSS-Kryptverfahren auf DVDs
- MMC-Befehle auch für ATAPI-Geräte definiert
- erlaubt gemeinsame Treiber für SCSI- und ATAPI-Geräte
- in aktuellen Geräten (etwa CD-Brenner) implementiert

SCSI: MMC vs. ATAPI

Annex B ATAPI Compliance (normative)

B.1. Introduction

This section describes the implementation of the MultiMedia Commands in ATAPI devices. The intent is to make the command sets highly compatible. It is desired that a common driver may be written to control both SCSI and ATAPI devices.

B.2. General

ATAPI devices implement a subset of SCSI behavior. Certain errors and conditions that exist in SCSI don't exist in ATAPI. In addition, certain terms are used in ATAPI instead of related SCSI terms. The mechanisms for transporting the commands, data, and status are unique to each transport. Addressing of units is also unique to each transport. MMC does not directly specify any of these mechanisms; the command and data layer definition may be layered on either transport.

B.2.1. Terms

B.2.1.1. Host - the ATAPI equivalent for the SCSI term "Initiator."

B.2.1.2. Device - the ATAPI equivalent for the SCSI term "Target" or "Logical Unit."

B.2.1.3. Command Packet - the ATAPI equivalent for the SCSI term "Command Descriptor Block."

B.2.2. Supported Block Sizes

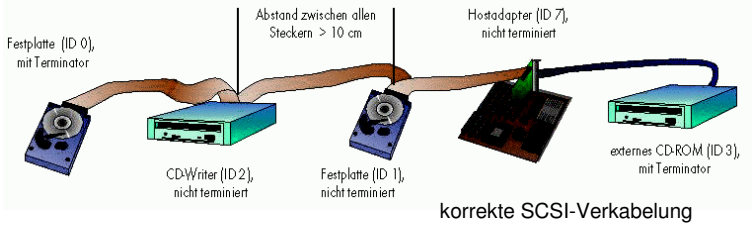
ATAPI does not use the block size in the mode select block descriptor. Instead, the block size shall be determined by the command. The READ command shall return 2048 bytes per block. The WRITE command shall send the number of bytes per block as determined by the WRITE PARAMETERS mode page. The READ CD and READ CD MSF commands shall return the number of bytes per block as specified by the command.

SCSI: MMC-Befehlssatz

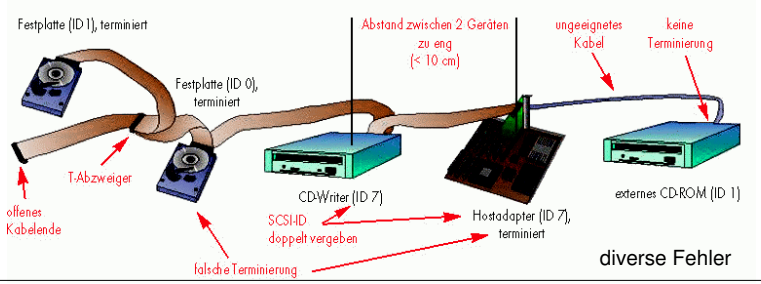
Command Name	Operation Code	MMC Type	Sub-clause
BLANK Command	A1h	O	6.1.1.
CLOSE TRACK/SESSION	5Bh	M	6.1.2.
FORMAT UNIT	04h	O	6.1.3.
LOAD/UNLOAD CD	A6h	O	6.1.5.
MECHANISM STATUS	BDh	M	6.1.6.
PAUSE/RESUME	4Bh	O	6.1.7.
PLAY AUDIO (10)	45h	A	6.1.8.
PLAY AUDIO (12)	A5h	A	6.1.9.
PLAY AUDIO MSF	47h	A	6.1.10.
READ BUFFER CAPACITY	5Ch	O	6.1.12.
READ CD	BEh	O	6.1.13.
READ CD MSF	B9h	M	6.1.14.
READ CD RECORDED CAPACITY	25h	M	6.1.15.
READ DISC INFORMATION	51h	M	6.1.16.
READ HEADER	44h	M	6.1.18.
READ MASTER CUE	59h	O	6.1.19.
READ SUB-CHANNEL	42h	M	6.1.21.
READ TOC/PMA/ATIP	43h	M	6.1.22.
READ TRACK INFORMATION	52h	O	6.1.23.
REPAIR TRACK	58h	O	
RESERVE TRACK	53h	M	6.1.28.
SCAN	BAh	O	6.1.30.
SEEK	2Bh	M	
SEND CUE SHEET	5Dh	O	6.1.31.
SEND OPC INFORMATION	54h	O	6.1.33.
SET CD SPEED	BBh	M	6.1.34.
STOP PLAY/SCAN	4Eh	O	
SYNCHRONIZE CACHE	35h	M	
WRITE (10)	2Ah	O	6.1.38.

- CD-Befehle:
  Load/Unload CD
  Play Audio (analog/dig.)
  Read CD (grabbing)
  Read Sub-Channel
  Read TOC / ...

SCSI: Verkabelung



korrekte SCSI-Verkabelung



diverse Fehler

Server: dimensionieren . . .

Ausgangslage und Aufgabe: [H&P, 530ff]

- Prozessor mit 500 MIPS, kostet $30.000
- Speicher, Busbreite 16 Byte, 100 ns Zykluszeit
- I/O-Bus mit 200 MB/s Bandbreite, Platz für 20 SCSI-2 Controller
- Betriebssystem benötigt 10.000 CPU-Befehle pro Platten-I/O
- SCSI-2 Busse, jeweils bis 20 MB/s, bis 15 Geräte (="SCSI String")
- SCSI-2 Controller a $1.500, mit 1 ms Latenzzeit pro I/O-Transfer
- Platten mit 2 GB oder 8 GB, Preis jeweils $0.25 pro MB
- beide Platten jeweils 7.200 rpm, 8 ms access time, 6 MB/s Transfer
- geforderte Speicherkapazität 200 GB
- mittlere Blockgröße für I/O-Transfers ist 16 KB

=> Performance mit kleinen/großen Platten? Kosten pro I/O-Transfer?
günstigste Konfiguration? wie viele Controller, welche Platten, usw.

Server: Grenzen durch CPU, Speicher, Bus

- IOPS = Anzahl I/O-Transfers pro Sekunde

$$\text{IOPS\_CPU} = \frac{500\ \text{MIPS}}{10.000\ \text{Befehle pro I/O}} = 50.000$$

$$\text{IOPS\_Speicher} = \frac{(1/100\ \text{ns}) \times 16\ \text{Byte}}{16\ \text{KB pro I/O-Transfer}} \sim 10.000$$

$$\text{IOPS\_Bus} = \frac{200\ \text{MB / s}}{16\ \text{KB pro I/O-Transfer}} \sim 12.500$$

=> Speicher limitiert auf maximal 10.000 IOPS

Server: Grenzen durch Controller und Platten

- Dauer eines SCSI-2 Transfers für 16 KB Daten:
- aber Controller benötigt 1 ms Overhead für den Transfer, also

$$t\_16KB = \frac{16\ \text{KB}}{20\ \text{MB / s}} = 0.8\ \text{ms}$$

$$\text{IOPS\_controller} = \frac{1}{(0.8\ \text{ms} + 1.0\ \text{ms})} \sim 556\ \text{IOPS}$$

- mittlere Dauer für Platten-I/O mit 16 KB Daten (zufällige Zugriffe):

$$t\_disk = 8\ \text{ms} + \frac{0.5}{7200\ \text{rpm}} + \frac{16\ \text{KB}}{6\ \text{MB / s}} = 8 + 4.2 + 2.7 = 14.9\ \text{ms}$$

$$\text{IOPS\_disk} = \frac{1}{14.9\ \text{ms}} \sim 67\ \text{IOPS}$$

Server: kleine oder große Platten

- 200 GB Kapazität: 25 8-GB Platten oder 100 2-GB Platten
- entsprechende Anzahl der IOPS:

IOPS_2GB = 100 x 67 = 6700
IOPS_8GB = 25 x 67 = 1675

- Mindestanzahl der Controller bei 15 Platten pro String

Strings_2GB = (100 / 15) = 7
Strings_8GB = (25 / 15) = 2

- Mindestanzahl der Controller, damit diese nicht der Flaschenhals?

Disks/String < 557 / 67 < 8

Strings_2GB = (100 / 8) = 12.5 = 13 (aufrunden)
Strings_8GB = (25 / 8) = 3.1 = 4 (aufrunden)

Server: Performance

Architekturen:

Typ	#Platten	#Controller	
2 GB	100	7 (min)	13 (opt)
8 GB	25	2 (min)	4 (opt)

Performance:

Platte	#SCSI	CPU	Speicher	Bus	Disks	Strings	IOPS	Kosten
8 GB	2	50.000	10.000	12.500	1675	**1112**	1112	$82.200
8 GB	4	50.000	10.000	12.500	**1675**	2224	1675	$87.200
2 GB	7	50.000	10.000	12.500	6700	**3892**	3892	$91.700
2 GB	13	50.000	10.000	12.500	**6700**	7228	6700	$100.700

- Server-Performance wird durch die Platten bzw. Controller limitiert (!)
- beste Performance mit vielen kleinen Platten und Controllern
- außerdem bestes Preis/IOPS-Verhältnis ($76, $52, $24, $15 pro IOPS)
- aber geringere Zuverlässigkeit (siehe RAID)

RAID: Motivation

Amdahl's Gesetz:
langsamste Komponente behindert Leistungssteigerungen

=> ausgewogenes Verhältnis CPU - Speicher - I/O nötig
=> CPU und Speicher skalieren mit der Halbleitertechnologie
=> aber wie kann die I/O-Leistung gesteigert werden?

RAID, "redundant array of inexpensive disks":

- Grundidee: viele kleine PC-Festplatten statt einer großen
- bedingt in damaliger (1985er) Festplattentechnologie:
  Großrechner-Festplatten vs. PC-Festplatten
- Zuverlässigkeit durch redundante Platten
- Wiederherstellung der Daten nach Plattenausfall

- ursprünglich: "independent disks"

Disks: RAID

"redundant array of inexpensive disks"

- bahnbrechende Untersuchung von Festplatten-Performance
- ursprünglich Analyse von Großrechner- und PC-Festplatten
- Ersetzen weniger großer durch viele kleine Festplatten
- Zuverlässigkeit des Gesamtsystems?

- diverse RAID-Varianten (=level)
- unterschiedliche Anzahl von Platten
- Strategien zur Verwendung von Nutz- und Reserveplatten
- Ausfallsicherheit, Hot-Plugging
- Optimierung auf Schreib- und/oder Leseperformance
- vielfache Anwendungen

- möglichst das Original lesen!

  [Patterson, Gibson, Katz: UCB report CS-98-391]

RAID: Ausgangsbasis (1987)

Characteristics	*IBM 3380*	*Fujitsu M2361A*	*Conners CP3100*	*3380 v. CP3100*	*2361 v. CP3100*
				(>1 means 3100 better)	
Disk diameter (inches)	14	10.5	3.5	4	3
Formatted Data Capacity (MB)	7500	600	100	.01	.2
Price/MB(controller incl.)	$18-$10	$20-$17	$10-$7	1-2.5	1.7-3
MTTF Rated (hours)	30,000	20,000	30,000	1	1.5
MTTF in practice (hours)	100,000	?	?	?	?
No. Actuators	4	1	1	.2	1
Maximum I/O's/second/Actuator	50	40	30	.6	.8
Typical I/O's/second/Actuator	30	24	20	.7	.8
Maximum I/O's/second/box	200	40	30	.2	.8
Typical I/O's/second/box	120	24	20	.2	.8
Transfer Rate (MB/sec)	3	2.5	1	.3	.4
Power/box (W)	6,600	640	10†	660	64
Volume (cu. ft.)	24	3.4	.03	800	11

Table I. *Comparison of IBM 3380 disk model AK4 for mainframe computers, the Fujitsu M2361A "Super Eagle" disk for minicomputers, and the Conners Peripherals CP 3100 disk for personal computers. By "Maximum I/O's/second" we mean the maximum number of average seeks and average rotates for a single sector access. Cost and reliability information on the 3380 comes from widespread experience [IBM 87] [Gawlick87] and the information on the Fujitsu from the manual [Fujitsu 87], while some numbers on the new CP3100 are based on speculation. The price per megabyte is given as a range to allow for different prices for volume discount and different mark-up practices of the vendors.*

†The 8 watt maximum power of the CP3100 was increased to 10 watts to allow for the inefficiency of ar external power supply (since the other drives contain their own power supplies).

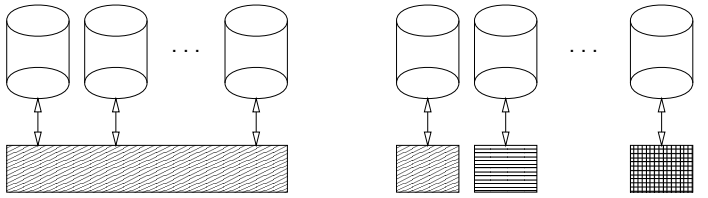
RAID: Kriterien

- Gesamtkapazität der Festplatte(n) — MByte
- maximale und typische Bandbreite — MByte/s
- maximale und typische Latenzzeiten — s
- Kosten, Volumen, Energieverbrauch — $, m³, W
- Zuverlässigkeit
  - MTTF, "mean time to failure" — s
  - MTTR, "mean time to repair" — s
  - MTTF_total = (MTTF_single / number_of_disks)

RAID-Konzept: viele parallele Platten

- höhere Gesamtkapazität, höhere Bandbreite
- Redundanz erhöht (!) die Zuverlässigkeit
- damalige Annahme: ca. 100 Platten, heute: typ. 5-10

RAID: Szenarien



welche Anwendungen benötigen hohe I/O-Leistung?

"scientific":	wenige, aber große Transfers
"database"	sehr viele kleine Transfers

RAID: Glossar

D	Gesamtanzahl der Platten
G	Anzahl der Daten- (=nutz) Platten pro Gruppe
NG	Anzahl der Gruppen
C	Anzahl der redundaten Check-Platten
rc	Verhältnis C/G
s	slowdown, typ. 1 < s < 2

RAID: Statistik

Annahmen zur Zuverlässigkeit der Platten:

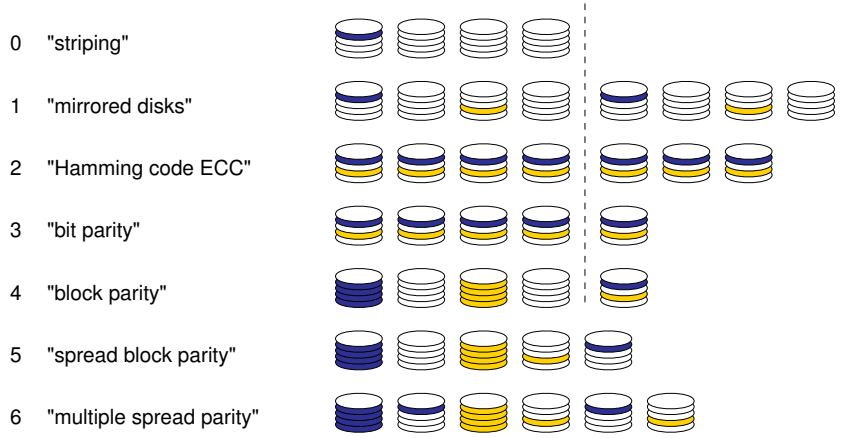
- Ausfälle sind zufällig, unabhängig, exponentialverteilt
- äußere Einflüsse (Sabotage, Stromausfall, ...) nicht berücksichtigt
- Controller ist robust

$$\text{MTTF\_group} = \frac{\text{MTTF\_disk}}{G + C} = \frac{1}{\text{probability of failure before repairing the dead disk}}$$

$$\text{P\_second\_failure} = \frac{\text{MTTR}}{\text{MTTF\_disk} / (G+C-1)}$$

$$\text{MTTF\_raid} = \frac{(\text{MTTF\_disk})^2}{(D+C*NG) * (G+C-1)*\text{MTTR}}$$

RAID: Level-Übersicht



RAID-0: Striping

- Aufteilen jedes (großen) Zugriffs in "Streifen"
  D = G, C = 0

- jede Platte verarbeitet Anteil 1/D
- jeder Zugriff benutzt alle Platten
- theoretisch D-fache Bandbreite für Lesen und Schreiben
- nur für genügend große Zugriffe

- aber keine Fehlertoleranz
- Zuverlässigkeit sinkt auf 1/D
- Einsatz nur für geringe Anzahl von Platten

- nicht im originalen "RAID paper" enthalten
- marktübliche "RAID-0" Kontroller verwalten zwei Platten

RAID-1: Mirroring

- Daten werden auf je zwei Platten "gespiegelt"
  G = 1, C = 1

- nutzt nur 50% der Gesamtkapazität der Platten
- jeder Schreibzugriff geht auf zwei Platten
- Schreibzugriff muß auf die jeweils langsamere Platten warten

- optimierte Version benutzt doppelten Kontroller
- erlaubt doppelte Bandbreite beim Lesen

- kein komplexer Controller notwendig
- ineffizient, aber sehr zuverlässig
  z.B. 500 Jahre MTTF

- keine besondere Marktbedeutung

RAID-2: Hamming Code ECC

- Hamming-Code zur Fehlerkorrektur jeder Gruppe von Platten
  z.B. (G=10, C=4) oder (G=25, C=5) usw.

- analog zur ECC-Fehlerkorrektur bei DRAMs
- Controller muß ECC berechnen und auswerten
- Aufteilung in Daten- und Check-Platten

- "große" Zugriffe laufen auf alle Platten einer Gruppe
- dabei volle Performance beim Lesen und Schreiben

- "kleine" Zugriffe kompliziert: gesamten Block lesen,
  ECC mit neuen Daten berechnen, gesamten Block schreiben
- daher sehr schlechte Gesamtperformance

- CRC-Code der einzelnen Platten unnötig
- sehr hohe Zuverlässigkeit, z.B. 50 Jahre MTTF mit G=10

RAID-3: Bit-Parität

- eine Platte mit Paritätscode pro Gruppe
  C=1

- Hamming-Code ermittelt, welche Platte Fehler aufweist
- dies liefert aber bereits der CRC jeder einzelnen Platte
- Paritätskode reicht aus, um den Fehler zu korrigieren

- weniger Checkdisks als RAID-2
- aber gleiches Performanceproblem für "kleine" Zugriffe
- jeder Schreibzugriff benutzt die Paritätsplatte

- weniger Platten als RAID-2, daher Preis/Leistung besser

- sehr hohe Zuverlässigkeit, z.B. 50 Jahre MTTF mit G=10

RAID-3: Vergleich Level 2 / 3

MTTF		Exceeds Useful Lifetime G=10 (820,000 hrs or >90 years)		G=25 (346,000 hrs or 40 years)	
Total Number of Disks		1.10D		1.04D	
Overhead Cost		10%		4%	
Useable Storage Capacity		91%		96%	
I/Os/Sec	Full RAID	Per Disk L3	L3/L2	Per Disk L3	L3/L2
(vs. Single Disk)					
Large Reads/sec	D/S	.91/S	127%	.96/S	112%
Large Writes/sec	D/S	.91/S	127%	.96/S	112%
Large R-M-W/sec	D/2S	.45/S	127%	.48/S	112%
Small Reads/sec	D/SG	.09/S	127%	.04/S	112%
Small Writes/sec	D/2SG	.05/S	127%	.02/S	112%
Small R-M-W/sec	D/2SG	.05/S	127%	.02/S	112%

Table IV. *Characteristics of a Level 3 RAID. The L3/L2 column gives the % performance of L3 in terms of L2 (>100% means L3 is faster). The performance for the full systems is the same in RAID levels 2 and 3, but since their are fewer check disks the performance per disk improves. Once again if the disks in a group are synchronized, then S = 1.*

RAID-4: Block-Parität

- eine Platte mit Paritätscode pro Gruppe
  C=1

- einzelner Datenblock wird auf eine einzelne Platte geschrieben
- Parität des Blocks auf die Paritätsplatte
- Paritätskode reicht aus, um den Fehler zu korrigieren

- gleiche Anzahl Platten wie RAID-3
- aber andere Organisation
- Lesezugriffe parallel ausführbar
- Schreibzugriffe parallel auf Datenplatten ausführbar
- aber Flaschenhals Paritätsplatte

- sehr hohe Zuverlässigkeit, z.B. 50 Jahre MTTF mit G=10

RAID-5: verteilte Parität

- Paritätscode auf alle Platten einer Gruppe verteilt
  C=1

- einzelner Datenblock wird auf eine einzelne Platte geschrieben
- Parität des Blocks auf die zugehörige Paritätsplatte
- Paritätskode reicht aus, um den Fehler zu korrigieren

- gleiche Anzahl Platten wie RAID-3
- aber effizienteste Organisation:
- Lesezugriffe parallel ausführbar
- Schreibzugriffe weitgehend parallel ausführbar

- attraktivste Variante, erfordert aber komplexen Controller
- hohe Zuverlässigkeit, z.B. 50 Jahre MTTF mit G=10

RAID-6: unabhängige, verteilte Parität

- mehrfacher Paritätscode auf alle Platten einer Gruppe verteilt
  C=2, 3, ...

- einzelner Datenblock wird auf eine einzelne Platte geschrieben
- Parität des Blocks auf die zugehörigen Paritätsplatten
- diverse Code-Varianten möglich

- ähnlich wie RAID-5
- aber bessere Fehlererkennung/korrektur
- Lesezugriffe parallel ausführbar
- Schreibzugriffe weitgehend parallel ausführbar
- noch komplexerer Controller als RAID-5

- nicht im originalen RAID-Paper erwähnt

RAID: vs. single disks (1987)

Characteristics	*RAID 5L (100,10) (CP3100)*	*SLED (IBM 3380)*	*RAID v. SLED (>1 better for RAID)*	*RAID 5L (10,10) (CP3100)*	*SLED (Fujitsu M2361A)*	*RAID v. SLED (>1 better for RAID)*
Formatted Data Capacity (MB)	10,000	7,500	1.33	1,000	600	1.67
Price/MB (controller incl.)	$11-$8	$18-$10	2.2-.9	$11-$8	$20-$17	2.5-1.5
Rated MTTF (hours)	820,000	30,000	27.3	8,200,000	20,000	410
MTTF in practice (hours)	?	100,000	?	?	?	?
No. Actuators	110	4	22.5	11	1	11
Max I/O's/Actuator	30	50	.6	30	40	.8
Max Grouped RMW/box	1250	100	12.5	125	20	6.2
Max Individual RMW/box	825	100	8.2	83	20	4.2
Typ I/O's/Actuator	20	30	.7	20	24	.8
Typ Grouped RMW/box	833	60	13.9	83	12	6.9
Typ Individual RMW/box	550	60	9.2	55	12	4.6
Volume/Box (cubic feet)	10	24	2.4	1	3.4	3.4
Power/box (W)	1100	6,600	6.0	110	640	5.8
Minimum Expansion Size (MB)	100-1000	7,500	7.5-75	100-1000	600	0.6-6

Table VII. *Comparison of IBM 3380 disk model AK4 to Level 5 RAID using 100 Conners & Associates CP 3100s disks and a group size of 10 and a comparison of the Fujitsu M2361A "Super Eagle" to a level 5 RAID using 10 inexpensive data disks with a group size of 10. Numbers greater than 1 in the comparison columns favor the RAID.*

RAID: Vergleich (1987)

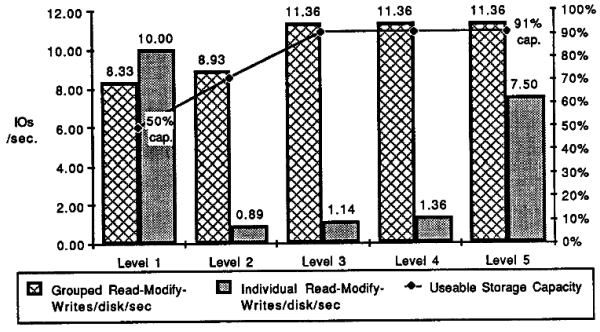
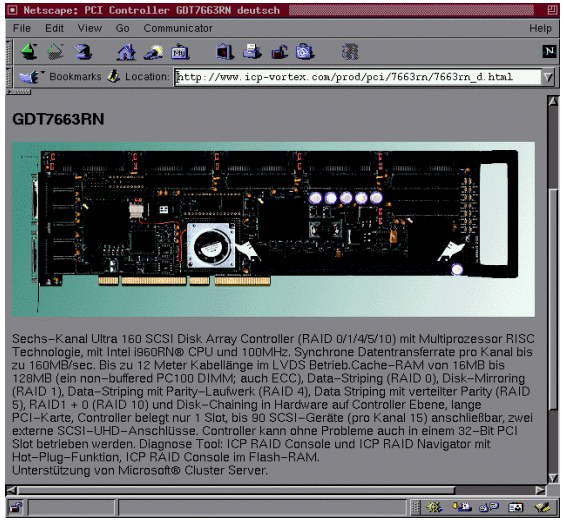


Figure 5. *Plot of Large (Grouped) and Small (Individual) Read-Modify-Writes per second per disk and useable storage capacity for all five levels of RAID (D=100, G=10, I/O=30/sec, S=1.2). To scale performance to other speed disks, simply multiply these numbers by the ratio to 30 I/O's/sec.*

- Level-1 schnell, sicher, teuer
- Level-5 der beste Kompromiss

RAID: Beispiel für einen Controller



Disks: Filecache

"Filecache"

- Plattenzugriffe deutlich langsamer als Speicherzugriffe
- häufig benutzte Daten (Dateien) im Hauptspeicher halten

=> Teil des Hauptspeichers als Filecache reservieren

- aber Filecache reduziert nutzbaren Hauptspeicher
- wo liegt das Optimum?

- nutzungsabhängig, single/multi user, workstation/server
- verschiedene Betriebssystemstrategien
- z.B. Windows 95 vs. Windows NT

- im folgenden einige Beispiele aus H&P

Disks: Filecache: Performance

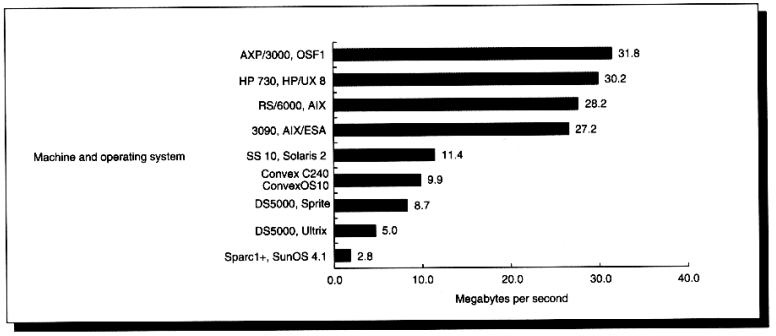


FIGURE 6.38 File cache performance for machines in 6.35. This plot is for 32-KB reads with the number of bytes touched limited to fit within the file cache of each system. Figure 6.36 (page 541) shows the size of the file caches that achieve this performance. (See the caption of Figure 6.36 for details on measurements.)

Disks: Filecache

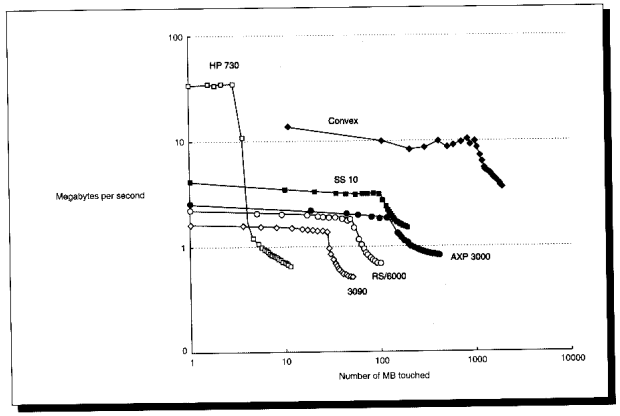


FIGURE 6.26 Performance versus megabytes touched for several workstations and mainframes (see section 6.8). Note the log-log scale. These results use the nominal values selected by the self-scaling benchmark. For example, 50% of accesses are reads and 50% are writes. The primary difference between the systems is the average access size of 120 KB for the Convex; adjusting for a common access size would halve Convex performance but make little change to the other lines in this plot.

[Hennessy & Patterson]

Disks: Filecache: Size

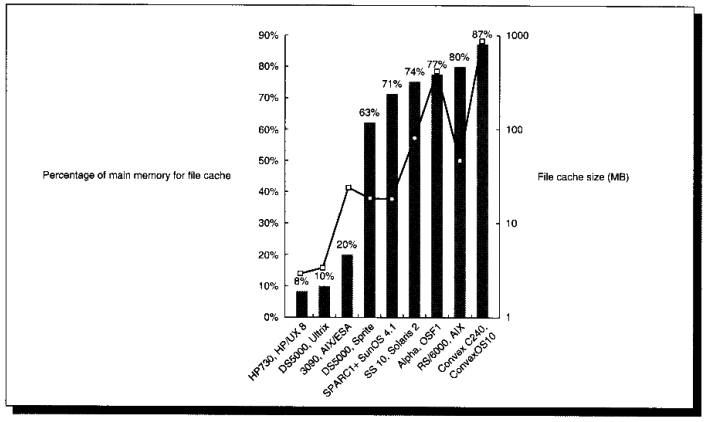


FIGURE 6.39 File cache size. The bar graph shows the maximum percentage of main memory for the file cache, while the line graph shows the maximum size in megabytes, using the log scale on the right. Thus the HP 730 HP/UX version 8 uses only 8% of its 32-MB main memory for its file cache, or just 2.7 MB, and the Convex C240 uses 87% of its 1024-MB main memory, or 890 MB, for its file cache.

Disks: Filecache: Read vs. Write

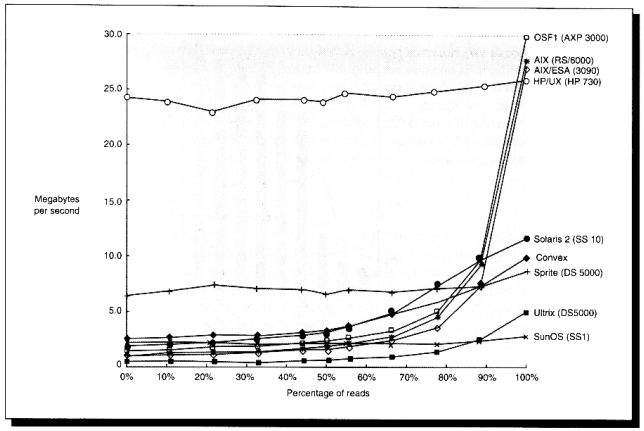


FIGURE 6.40 File cache performance versus read percentage. 0% reads means 100% writes. These accesses all fit within the file caches of the respective machines. Note that the high performance of the file caches of the AXP/3000, RS/6000, and 3090 are only evident for workloads with ≥ 90% reads. Access sizes are 32 KB. (See the caption of Figure 6.36 for details on measurements.)

Leerseite

Leerseite

Leerseite