7. File I/O

AlDaBi Praktikum

David Weese
WS 2010/11

Inhalt

- Externspeicher
- Externspeichermodelle
- I/O-Interfaces in C++

- Bemerkungen zur P-Aufgabe

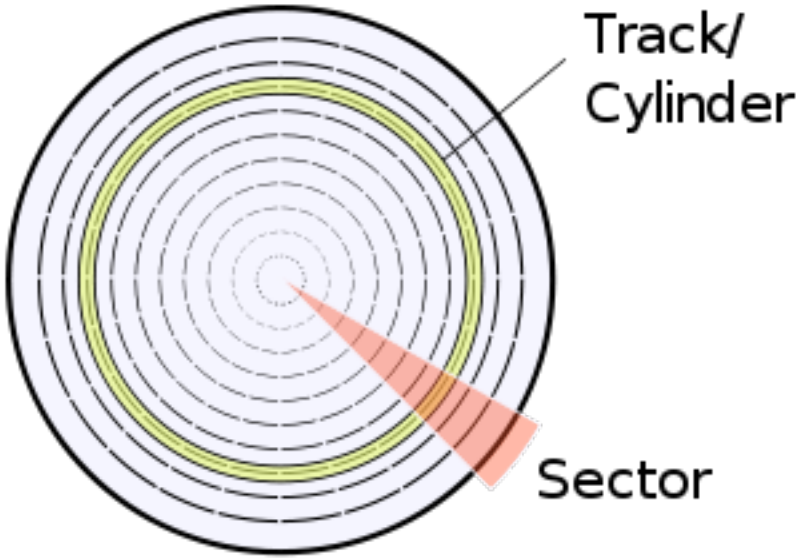
EXTERNSPEICHER

Aufbau einer Festplatte (HDD)

- Typische Bestandteile einer Festplatte:
  - Schnell rotierende **Spindel** (spindle), typ. 3.000-15.000 U/min
  - An der Spindel befestigte flache ferromagnetische **Scheiben** (platter)
  - Beweglicher **Arm** an dessen Ende die Köpfe montiert sind
  - Schreib-Lese-**Köpfe** (head) für jede Seite einer Scheibe

Datenanordnung

- Gespeicherte Daten liegen auf den Scheibenoberflächen und werden von den Köpfen gelesen/geschrieben (head)
- Scheibenoberfläche ist in konzentrische Spuren (track/cylinder) unterteilt
- Spur besteht aus Sektoren (sector)

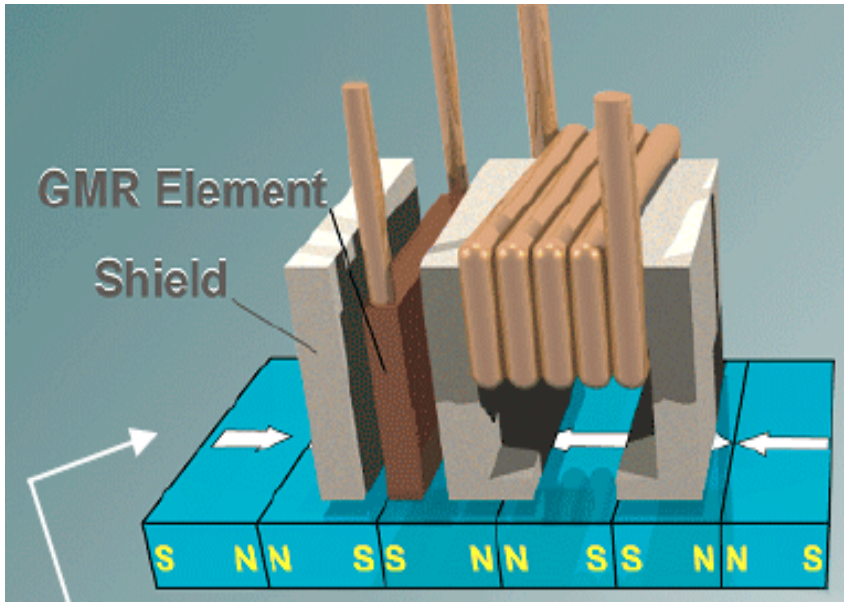


- Typische Sektorgröße:
  - 512 Byte

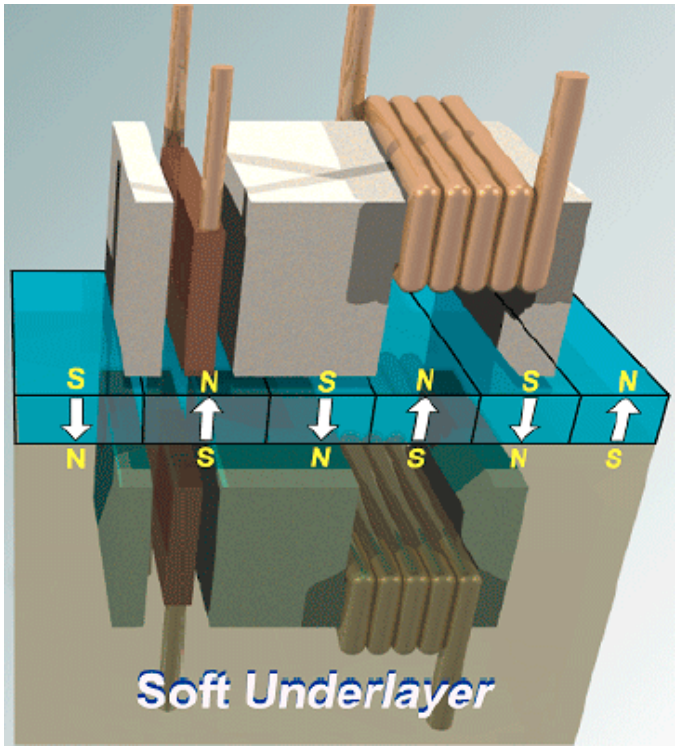
- Adresse eines Sektors:
  - CHS-Tripel
    (**C**ylinder, **H**ead, **S**ector)

Schreib-Lese-Kopf

- Ausrichtung der magnetischen Domänen kodiert Daten (RLL, MFM)
- Magnetische Domänen werden entweder entlang der Kopfbewegung (longitudinal) oder senkrecht dazu (perpendicular) ausgerichtet



Longitudinal (vor 2006)

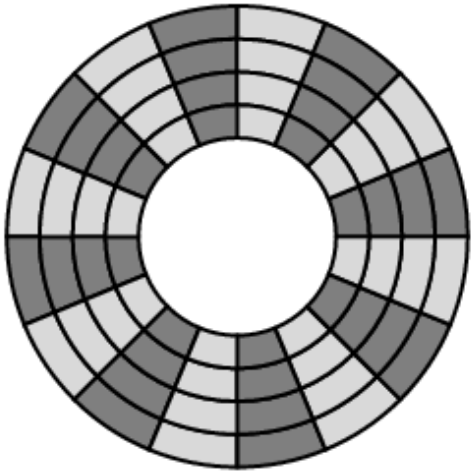


Perpendicular (aktuell)

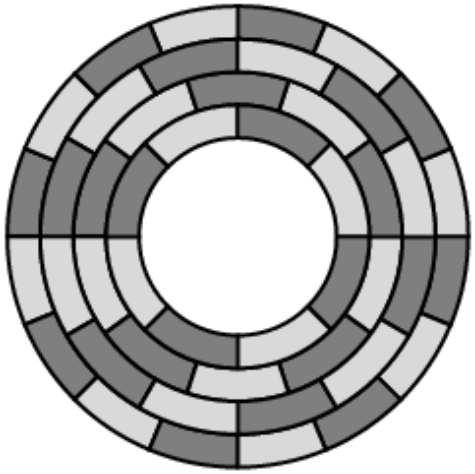
Datendichte

- Festplatten rotieren mit konstanter Geschwindigkeit
  - nach außen nimmt die Oberflächengeschwindigkeit der Köpfe zu
  - Haben alle Sektoren das gleiche Bogenmaß (CAV) nimmt die Datendichte nach außen ab
  - Aktuelle Festplatten wählen möglichst gleichgroße Sektoren (MZR)

Constant **A**ngular **V**elocity



Multiple **Z**one **R**ecording



Performanz

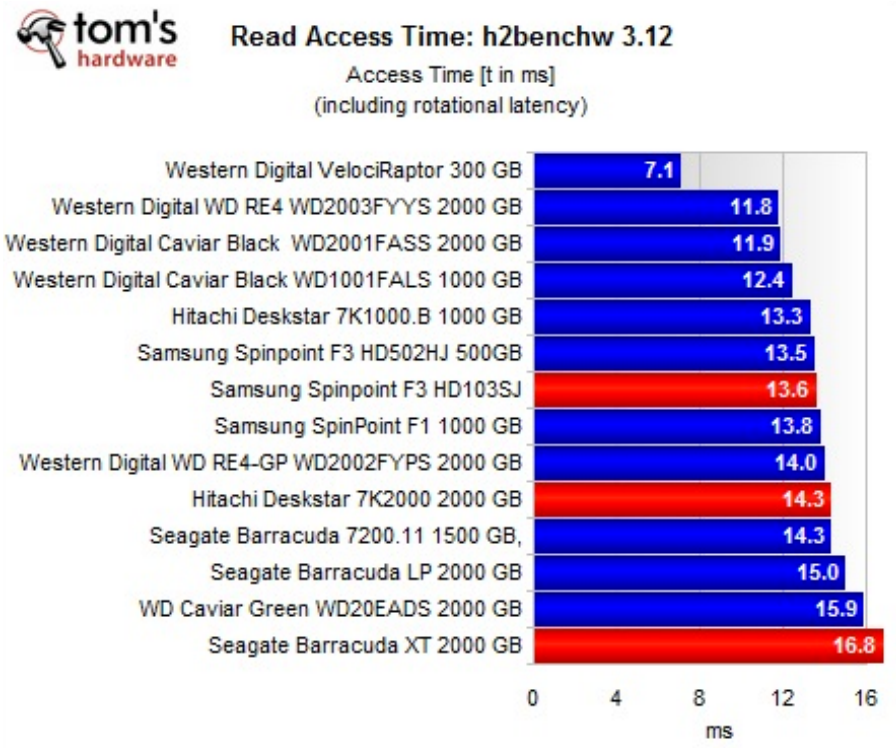
- **Zugriffszeit** ist die Summe aus:
  - **Seek Time**
    - Bewegung des Armes (Beschleunigung, Max. Geschwindigkeit, Abbremsen)
    - Langsam (Millisekunden)
  - **Rotational Delay**
    - Warten bis rotierender Sektor den Kopf erreicht hat
    - Durchschnitt: Zeit für eine halbe Plattenrotation
    - Langsam (Millisekunden)
  - **Controller Overhead**
    - Zeit die der Festplattencontroller zum Interpretieren und Koordinieren des Kommandos braucht
    - Schnell

- **Durchsatz:**
  - Hängt ab von Datendichte und Oberflächengeschwindigkeit
  - Steigt von inneren zu äußeren Spuren (bei Multiple Zone Recording)

Zugriffszeiten

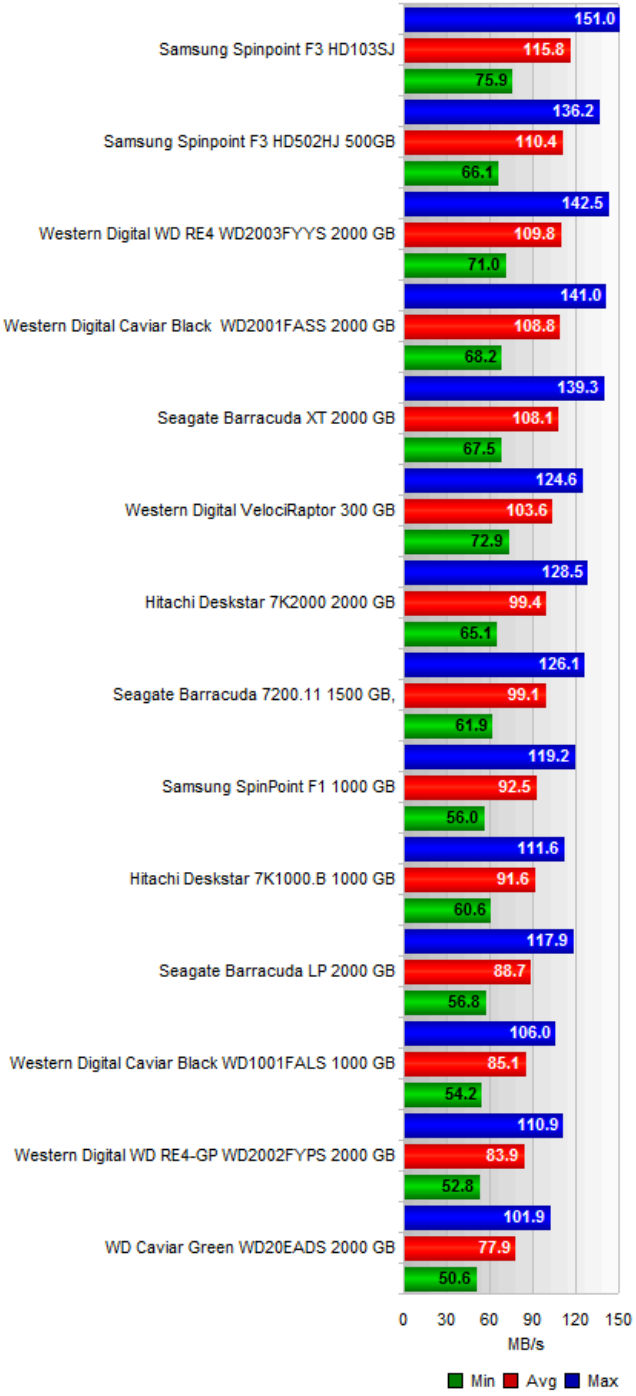
- Durchschnittliche Zugriffszeiten (2010):
  - Zwischen 7 und 17ms
  - Tendenziell haben Platten mit größerer Kapazität höhere Zugriffszeiten

- Zugriffszeiten im Vergleich:
  - RAM 10-20ns (Faktor 10^6)
  - L1-Cache 1ns (Faktor 10^7)



Durchsätze

- Durchschnittlicher Durchsatz (2010):
  - Zwischen 50 und **150 MB/s**
  - Variiert mit Rotationsgeschwindigkeit und Entfernung der Daten vom Mittelpunkt
- Zugriffszeiten im Vergleich:
  - RAM 1-5 GB/s (nur noch Faktor 30)
- Durchsatz bei byteweisen wahlfreien Zugriffen:
  - 1 Byte/Zugriffszeit = 1 Byte/7ms = **143 B/s**
- **Wahlfreie Zugriffe vermeiden!!!**
- **Benutze Blocktransfers**



Speichermodelle

- Schlussfolgerungen:
  1. Wahlfreie Zugriffe entsprechen den Cache-Misses im Hauptspeicher
     - Techniken um Cache-Misses zu reduzieren lassen sich übertragen auf Externspeicher
     - Wahlfreier Zugriff ist im Verhältnis aber **10.000 mal** langsamer als ein Cache-Miss
  2. Algorithmen die im Hauptspeicher effizient sind lassen sich i.d.R. nicht effizient 1-zu-1 auf Externspeicher übertragen
  3. Die asymptotische Laufzeit (im RAM-Modell) sagt wenig über die tatsächliche Laufzeit mit Externspeicher aus
     - Es bedarf eines Speichermodells, das die Eigenschaften von Festplatten widerspiegelt

EXTERNSPEICHERMODELLE

Aggrawal und Vitter

- Erstes Modell Aggrawal und Vitter 1988
  - Computer mit Hauptspeicher der Größe **M**
  - Festplatte mit **P** Schreib-Lese-Köpfen
  - Festplattendaten können nur in Blöcken der Größe **B** gelesen und geschrieben werden

- Gezählt wird die Zahl der I/O-Zugriffe
  - Das parallele Lesen von Daten der Größe P·B zählt als 1 Zugriff

- Nachteile dieses Modells:
  - In aktuellen Platten sind Köpfe nicht unabhängig voneinander beweglich
  - Köpfe haben auch nur Zugriff auf eine (nicht jede) Oberfläche

[1] Aggarwal, A.; Vitter, J.S.: **The Input/Output Complexity of Sorting and Related Problems.** In: Communications of the ACM, 1988

Vitter und Shriver

- Verbesserung durch Vitter und Shriver 1994
  - Computer mit Hauptspeicher der Größe **M** und **P** Prozessoren
  - **D** Festplatten mit jeweils einem Schreib-Lese-Kopf
  - Festplattendaten können nur in Blöcken der Größe **B** gelesen und geschrieben werden

- Nachteile beider Modelle [1,2]:
  - Es wird nicht unterschieden zwischen wahlfreien (random) und sequenziellen (bulk) I/O-Zugriffen
  - Wahlfreie Zugriffe sind in der Praxis sehr viel langsamer als sequenzielle

[2] Vitter, J.S.; Shriver, E.A.M.: **Algorithms for Parallel Memory I: Two-Level Memories.** In: Algorithmica, 1994, vol. 12 (2) pp. 110-147

Farach et al.

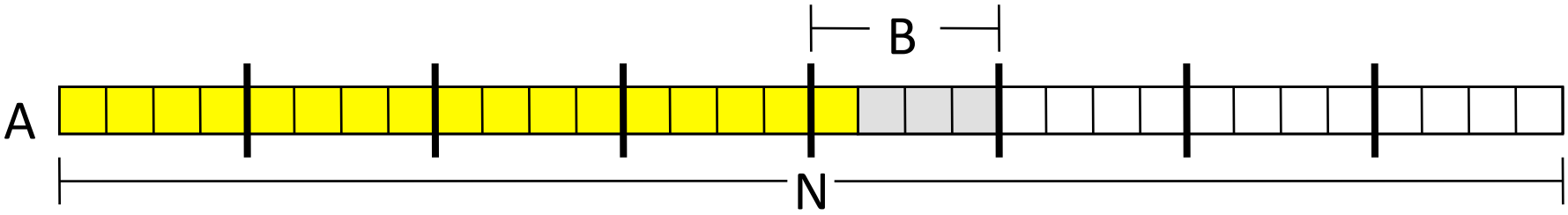
- Verbesserung durch Farach et al. 1998
  - Schließt sequentielle Zugriffe in Analyse ein
  - ≥M/B zusammenhängende Zugriffe heißen *sequentiell*

[3] Farach, M.; Ferragina, P.; Muthukrishnan, S.: **Overcoming the Memory Bottleneck in Suffix Tree Construction.** In: FOCS '98, Washington, DC, USA: IEEE Computer Society, 1998. – ISBN 0–8186–9172–7, S. 174

Linearer Scan

- Sequentielle Zugriffe auf ein Feld der Größe N
  - Beispiel:

```
int sum = 0;
for (int i = 0; i < N; ++i)
    sum += A[i];
```

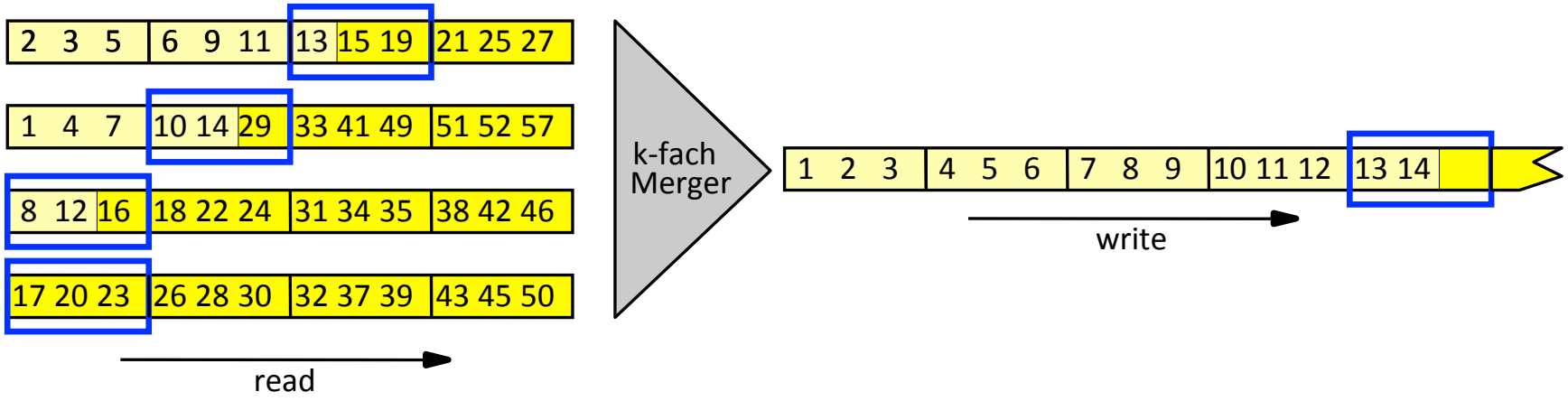


O(N/B) I/O-Zugriffe (bulk)

* Abbildungen von Gerth Stølting Brodal

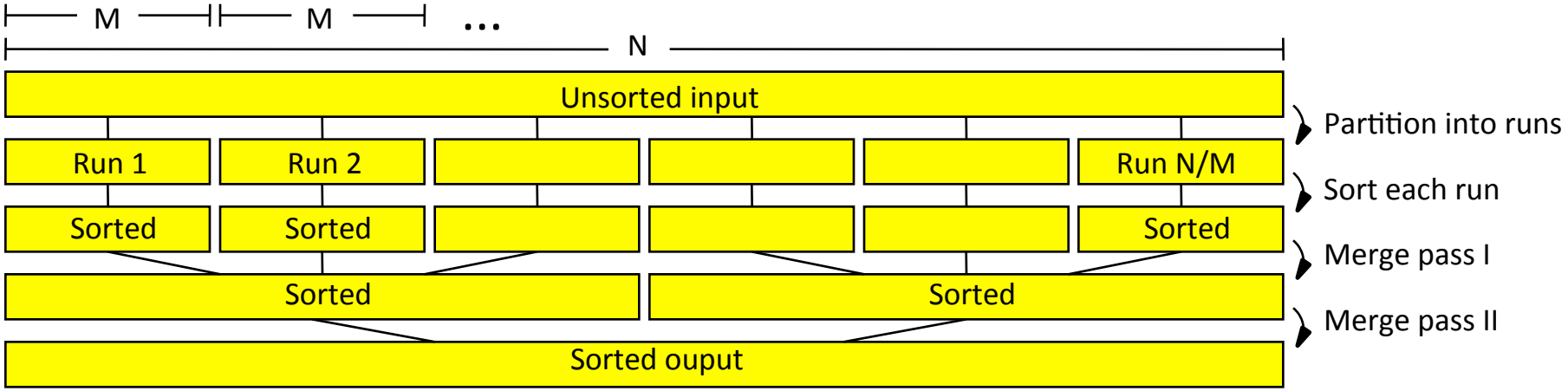
k-fach Merge

- Merge von k Sequenzen mit insgesamt N Elementen braucht $O(N/B)$ I/Os für $k + 1 \leq M/B$



Externes Sortieren

- θ(M/B)-faches MergeSort erreicht optimale *O*(Sort(N)=O(N/B·$\log_{M/B}$(N/B)) I/Os
- Aggarwal and Vitter 1988



I/O-INTERFACES IN C++

Verfügbare Interfaces

- Platformunabhängig:
  - C Standard Library (stdio.h)
  - C++ Standard Library File Streams (fstreams)

- Platformabhängig:
  - POSIX Files (fcntl.h, aio.h)
  - Windows CRT* (io.h)
  - Windows SDK (windows.h)
  - Memory Mapped Files (sys/mman.h oder windows.h)

* C Run-Time Libraries

POSIX

- POSIX (Portable Operating System Interface) [1]
  - Standard für Betriebssysteme
  - Inhalt:
    - Basis-Defintionen (bspw. \n entspricht newline)
    - Systemschnittstellen (C-Header, bspw. stdio.h)
      - Threads, Network, **Files**, Interprocess Communication, ..., **Memory Mapped Files**
    - Vorgaben für Shells und Programme

- Weitestgehend POSIX-kompatibel sind:
  - Linux
  - xBSD
  - Mac OS X
  - Solaris

[1] **POSIX.1:2008**, http://pubs.opengroup.org/onlinepubs/9699919799

Öffnen einer Datei

- Funktion zum Öffnen benötigt:
  - **Dateiname**
    - Kann Pfad enthalten, Slashes beachten (Windows: \\ und Linux, Mac: /)
  - **Flags**
    - Soll Datei erzeugt, erweitert, nur gelesen, nur geschrieben werden?
    - Beispiel (POSIX): O_CREAT, O_APPEND, O_RDONLY, O_WRONLY, O_RDWR, ...
    - Beispiel (C Lib): “w+“, “a+“, “r“, “w“, “r+“
  - **Rechte**
    - Wenn Datei erzeugt wird, welche Zugriffsrechte soll sie haben (Bsp.: 755)?

- Rückgabe:
  - **Dateihandle**
    - Wird von allen folgenden Dateioperationen benötigt
    - Hat bestimmten Wert bei Fehler (bspw. -1)
    - Fehlercode kann abgefragt werden (bspw.: errno)

I/O Zugriff

- Blockweise
  - Liest/Schreibt einen zusammenhängenden Teil der Datei direkt in den/ aus dem Speicher
  - Synchron/Asynchron
  - Datei kann auch direkt in den Speicher eingeblendet werden (Memory Mapping)
  - Schnell und damit als externer Speicher geeignet

- Zeichenweise
  - Liest/Schreibt einzelne Zeichen/Zeilen der Datei (nur synchron)
  - Teilweise mit automatischer Konversion der Zeilenendungen
  - Langsam und gedacht zum Parsen/Ausgeben von Dateiformaten

Synchroner Dateizugriff

- Synchroner Dateizugriff
  - Lese- und Schreiboperationen sind **blockierend**
  - Aufrufer-Thread pausiert (suspend) bis zum Abschluss der Operation

- Dateien
  - Zu einer geöffneten Datei gehört ein **Dateizeiger** = Position innerhalb der Datei
  - Operationen werden ab dem Dateizeiger ausgeführt
  - Dateizeiger kann verschoben (seek) oder abgefragt werden (tell)

- Streams
  - Streams haben **keinen Dateizeiger** (seek und tell schlagen fehl)
  - Beispiele: cin, cout, cerr

Synchroner Dateizugriff

C Standard Library

```c
#include <stdio.h>

FILE* fp = fopen(path, type);
fseek  (fp, offset, origin);
fread  (buf, num, len, fp);
fwrite (buf, num, len, fp);
fclose (fp);
```

C++ Standard Library

```c
#include <fstream>

std::fstream fs(path, flags);
fs.seekg (fp, offset, origin);
fs.read  (buf, size);
fs.write (buf, size);
```

POSIX

```c
#include <unistd.h>

int fd = open(path, flags);
lseek (fd, offset, origin);
read  (fd, buf, size);
write (fd, buf, size);
close (fd);
```

Windows CRT

```c
#include <io.h>

int fd = _open(path, flags);
_lseek (fd, offset, origin);
_read  (fd, buf, size);
_write (fd, buf, size);
_close (fd);
```

Asynchroner Dateizugriff

- Asynchroner Dateizugriff
  - Lese- und Schreiboperationen sind **nicht-blockierend**
  - Parameter der asynchronen Operation wird in `struct` gespeichert
  - Asynchrones `read`/`write` benötigt Parameter-Struktur und kehrt sofort zurück
  - Es gibt Funktionen zum ...
    - Abfragen des Zustands (läuft noch/fertig/fehlerhaft) einer Operation
    - Blockierenden Warten auf Abschluss einer Operation

- Operationen können in beliebiger Reihenfolge/parallel abgearbeitet werden
- Dateizeiger wird ignoriert
  - Startposition steht in Parameter-Struktur

Asynchroner Dateizugriff in C

POSIX [1]

```c
#include <aio.h>
#include <fcntl.h>
#include <errno.h>

int fd = open("ex.txt", O_RDONLY, 0);

aiocb cb;
memset(&cb, 0, sizeof(aiocb));
cb.aio_fildes = fd;
cb.aio_offset = 0;
cb.aio_buf    = buf;
cb.aio_nbytes = size;
aio_read(&cb);

...



if (aio_error(&cb) == EINPROGRESS)
     printf("Still reading...\n");
```

Windows SDK [2]

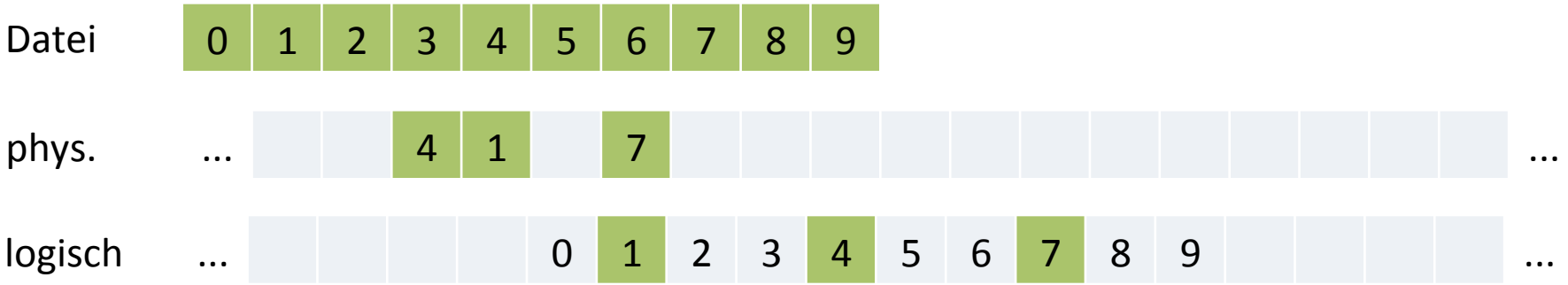
```c
#include <windows.h>

HANDLE fh = CreateFileA("ex.txt",
  GENERIC_READ, …, FILE_FLAG_OVERLAPPED,
  NULL);

OVERLAPPED cb;
cb.Offset     = 0;
cb.OffsetHigh = 0;
cb.hEvent     = NULL;


ReadFile(fh, buf, size, NULL, &cb);

...

LPWORD xmit_size;
GetOverlappedResult(fh, &cb, &xmit_size,
  FALSE);

if (GetLastError(&cb) == ERROR_IO_PENDING)
     printf("Still reading...\n");
```


[1] **Understanding the Linux Kernel 3rd Edition,** Kap. 16.4., http://book.opensourceproject.org.cn/kernel/kernel3rd
[2] **MSDN – Synchronous and Asynchronous I/O,** http://msdn.microsoft.com/en-us/library/aa365683(v=VS.85).aspx

Memory Mapped Files

- Was sind Memory Mapped Files?
  - Dateien, die in den (logischen) Adressraum abgebildet werden
  - Erst bei Zugriff auf den Speicherbereich wird der entsprechende Block in den physikalischen Speicher geladen
  - Lange nicht benutzte, veränderte Blöcke werden zurückgeschrieben (analog zu Caches)

- Die Dateien können größer sein als physikalische Speicher
  - Nicht aber größer als logischer Adressraum (4GB auf 32bit Systemen)



Memory Mapped Files (II)

- Memory Mapping erstellen und beenden:
  - POSIX:

```
void *addr = mmap(NULL, fileSize, …, fileHandle, 0);
…
munmap(addr, fileHandle);
```

  - Windows:

```
HANDLE handle = CreateFileMapping(fileHandle, …);
void *addr = MapViewOfFile(handle, …);
…
UnmapViewOfFile(addr);
CloseHandle(handle);
```

- `addr` zeigt auf das erste Byte der Datei
  - kann in `char*` Zeiger konvertiert und wie ein Feld benutzt werden

Fallstricke beim Blockweisen Zugriff

- Größe des Pufferspeicher entsprechend wählen
- Puffer muss zusammenhängend sein (bspw.: Feld, std::vector)
  - nicht zusammenhängend sind std::list, std::map, ...
- Für asynchrones I/O muss auf manchen Plattformen Pufferadresse durch Sektorgröße teilbar sein
- Klassen können größer sein als die Summe ihrer Elemente (Padding, P-VL3)
  - Notwendig damit bspw. ints an durch 4 teilbaren Adressen beginnen
  - Verschenkter Platz im Externspeicher
  - Daher: Elemente umordnen, Klassen packen

GCC

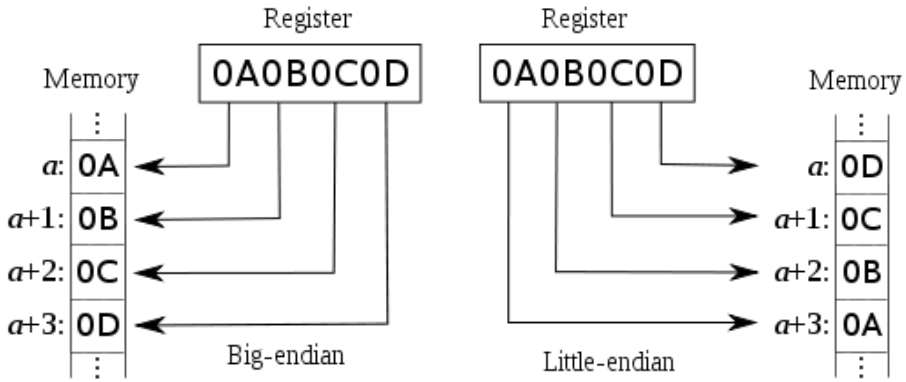
```
struct test_t {
  int  a;
  char b;
  int  c;
} __attribute__((__packed__));
```

Visual Studio

```
#pragma pack(push,1)
struct test_t {
  int  a;
  char b;
  int  c;
};
#pragma pack(pop)
```

Fallstricke beim Blockweisen Zugriff (II)

- Blockweise geschriebene Datenstrukturen haben das Format, das sie bei der Ausgabe im Speicher hatten
  - Ohne weiteres nicht plattformunabhängig (Sparc Solaris ↯ x86 Linux)
  - Größe der Datentypen ist plattformabhängig
    - long ist entweder 32 oder 64 bit breit je nach Plattform
    - Abhilfe schaffen die typedefs in stdint.h: uint8_t, int32_t, uint64_t, ...
  - Paddinganforderungen sind plattformabhängig
    - Umordnen, Klassen packen
  - Endianess ist plattformabhängig
    - Festlegen auf eine Endianess (bspw. Network Byte Order = Big Endian) und auf der anderen Plattform vor Schreib- und nach Lesezugriff um drehen (htonl, ntohs, ...)



Zeichenweiser I/O Zugriff

- C++ File Streams:
  - ifstream (nur lesen)
  - ofstream (nur schreiben)
  - fstream (beides)
- Sind abgeleitet von istream, ostream oder iostream
  - File Streams haben also mind. die Funktionalität von `cin` und `cout`
- File Stream öffnen:

```
std::fstream fs;
fs.open(path, flags);
…
fs.close();
```

- Oder öffnen im Konstruktor und schließen im Destruktor:

```
std::fstream fs(path, flags);
```

Einzelne Zeichen lesen/schreiben

- Ein Zeichen lesen:

```
char c;
c = fs.get();      // char lesen und weitergehen
c = fs.peek();     // char lesen aber nicht weitergehen
fs.unget();        // ein Zeichen zurückgehen
```

- Ein Zeichen schreiben:

```
fs.put(c);         // char schreiben und weitergehen
```

- Stream-Status abfragen:

```
bool b;
b = fs.is_open(); // Ist die Datei geöffnet?
b = fs.eof();     // Hat letzte Operation Streamende überschritten?
b = fs.good();    // Weder EOF noch Formatfehler?
```

Formatierte Ein-/Ausgabe

- ostream hat operator << für alle eingebauten Typen überladen:

```
int i = 42;
double d = 3.14;
fs << i;
fs << "  ";
fs << d;                        // Datei: “42  3.14“
```

- operator << gibt Referenz auf Stream zurück → Schachtelung möglich:

```
((fs << i) << "  ") << d;
fs << i << "  " << d;          // wird von links nach rechts ausgeführt
```

- operator >> (istream) überspringt Leerzeichen und liest Wert:

```
fs >> i;                // überspringe Leerzeichen, lies Ganzzahl
fs >> d;                // überspringe Leerzeichen, lies Gleitkommazahl

std::string s;
fs >> i >> s;           // lies Ganzzahl, lies String bis Leerzeichen
```

Zeilenende

- Lies String bis Zeilenende oder bis zu frei wählbarem Zeichen:

```
getline (fs, str);                  // lies ganze Zeile
getline (fs, str, ',');             // lies bis Komma
```

- Zeilenende schreiben:

```
fs << "Eine Zeile\n";                // eine Zeile schreiben
fs << "Eine Zeile" << std::endl;     // eine Zeile schreiben + flush
```

  - flush leert synchron Schreibpuffer in Datei (langsam)
- Zeilenende wird verschieden kodiert:
  - Windows benutzt CR+LF (`\r\n`)
  - Mac OS X benutzt LF (`\n`)
- IOStreams konvertieren Zeilenendungen implizit in \n
  - Konvertierung kann mit ios::binary abgeschaltet werden (schneller)

Stream Manipulatoren

- C++ bietet verschiedene Stream Manipulatoren mit versch. Aufgaben:
- Stream Manipulatoren (definiert in <iomanip>) beeinflussen:
  - Setzen der Breite und des Füllzeichens von Feldern
    - cout << setw(6) << setfill(0) << 3.2; // 0003.2
  - Setzen der angezeigten Präzision von Gleitkommazahlen
    - cout << setprecision(3) << 2.71828183; // 2.72
  - Setzen und Löschen von Formatierungsflags
  - Flushen von gepufferten Streams
    - cout << flush;
    - cout << endl; // entspricht cout << '\n' << endl;

Formatierungsflags

- ios::showpoint
  - Dezimalpunkt immer anzeigen
- ios::showpos
  - “+” vor positiven Zahlen anzeigen
- ios::basefield
  - Wählt die Zahlenbasis
    - ios::dec Dezimalzahlen
    - ios::oct Oktalzahlen
    - ios::hex Hexadezimalzahlen
- ios::floatfield
  - Art der Gleitkommadarstellung
    - ios::fixed feste Anzahl Stellen
    - ios::scientific wissenschaftliche Notation
- ios::adjustfield
  - Horizontale Ausrichtung
    - ios::left linksbündig
    - ios::right rechtsbündig
    - ios::internal Vorzeichen linksbündig, Wert rechtsbündig

BEMERKUNGEN ZUR P-AUFGABE

Hinweise zu Aufgabe 7

- DNA-Sequenzen in FASTA-Format einlesen, Reverskomplement bilden, in FASTA-Format rausschreiben
  - FASTA-Format:

```
>Sequenz 1
;Kommentarzeile A
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
>Sequenz 2
;Kommentarzeile B
;Kommentarzeile C
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
TTT
```

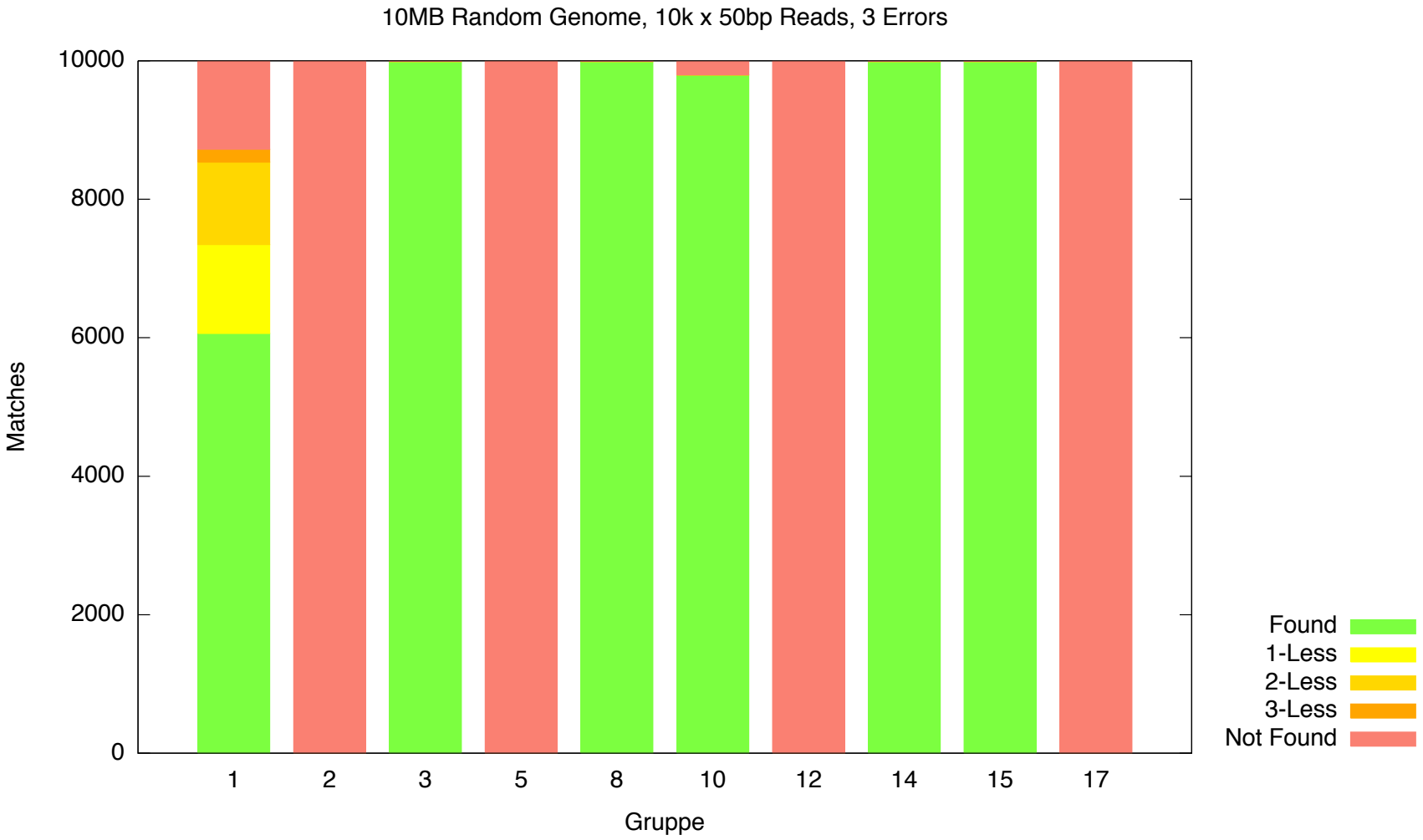
  - Reverskomplement
    - ACCGTA (Original) → TGGCAT (Komplement) → TACGGT (Reverskomplement)

Bemerkungen zu Aufgabe 5

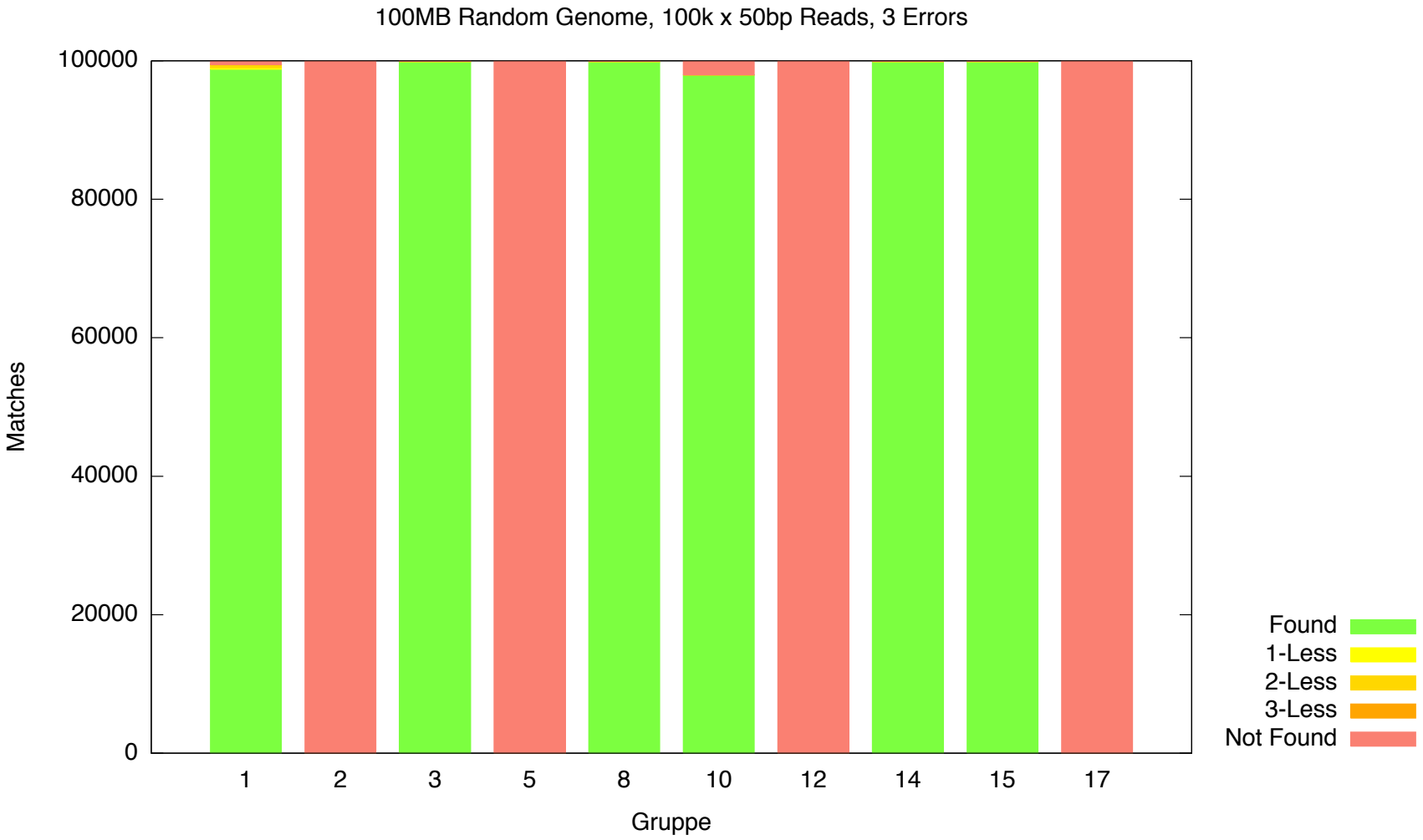
- Testdaten
  - Für den Benchmark wurden 10 und 100 Mbp zufällige DNA-Sequenz generiert
  - Daraus wurden jeweils zufällig 10.000 und 100.000 Reads der Länge 50bp ausgeschnitten
  - In die Reads wurden zufällig bis zu 3 Fehler implantiert
  - Ursprungsposition und Fehlerzahl jedes Reads wurden als **Seed-Match** notiert

- Auswertung
  - Jedes Programm wurde auf den beiden Datensätzen laufen gelassen
  - Maschine hatte 16 Cores, 72GB Hauptspeicher
  - Die ausgegeben Matches wurden mit den Seed-Matches verglichen
    - Zu jedem Seed-Match musste das Programm einen Match mit höchstens der angegeben Seed-Match-Fehlerzahl finden im Umkreis von 200bp (Blocktoleranz, einer pro Block reichte)
  - Gezählt wurden wieviele Seed-Matches ...
    - mit höchstens der Seed-Fehlerzahl gefunden wurden (**Found**)
    - mit einer Fehlerzahl gefunden wurden, die um i schlechter ist (**i-Less**)
    - gar nicht gefunden wurden (**Not Found**)

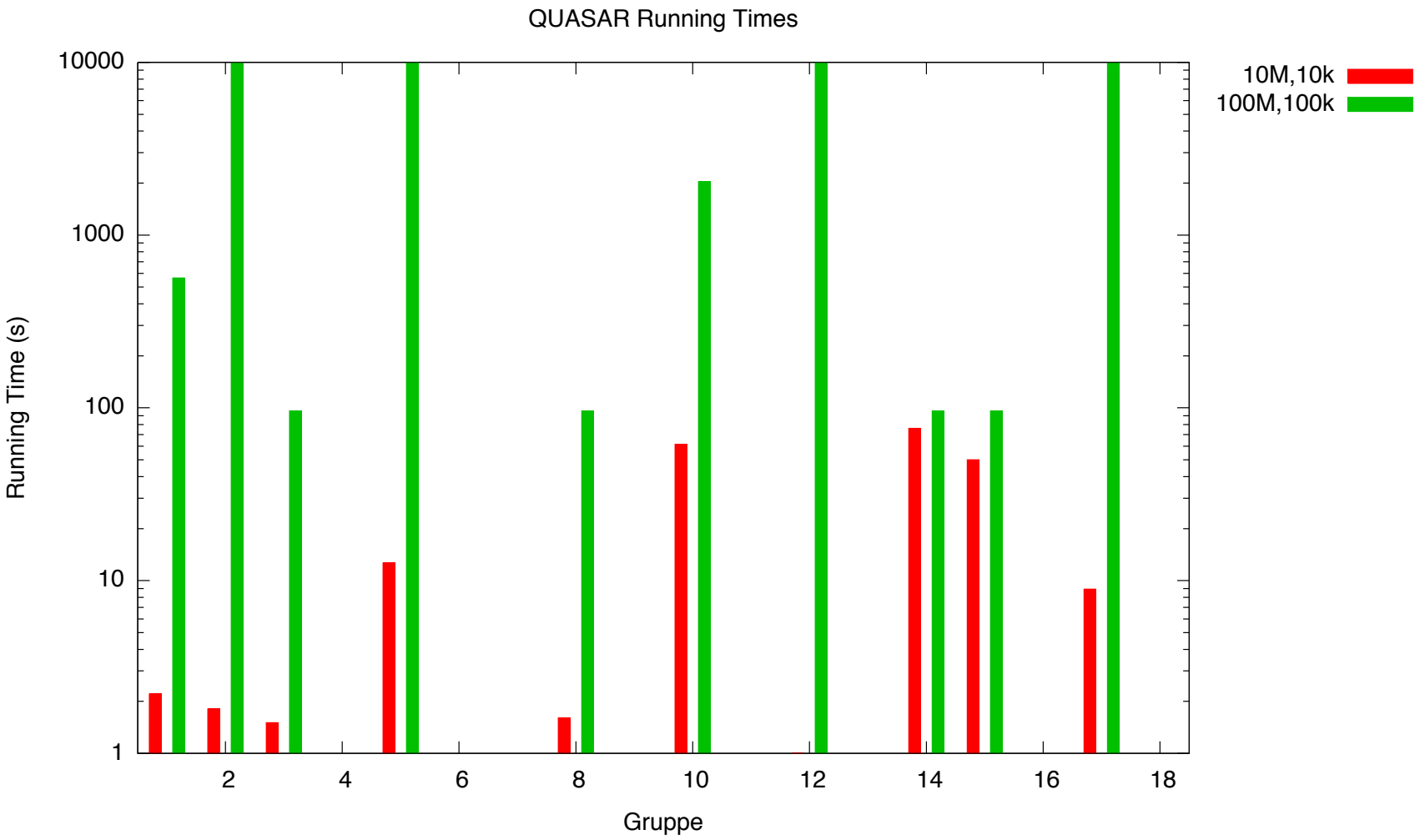
Sensitivität



Sensitivität (II)



Laufzeiten


QUASAR Running Times
Running Time (s)
10000
1000
100
10
1
2
4
6
8
10
12
14
16
18
Gruppe
10M,10k
100M,100k

Gewonnen hat Gruppe 8

Gruppe	Laufzeit (s)	Not-Found	Found	1-Less	2-Less	3-Less	4-Less
1	203,95	565	98792	210	405	28	0
2	36,18	100000	0	0	0	0	0
3	199,88	96	99862	42	0	0	0
5	81,38	100000	0	0	0	0	0
8	**114,26**	**96**	**99862**	**42**	**0**	**0**	**0**
10	410,77	2031	97969	0	0	0	0
12	8,21	100000	0	0	0	0	0
14	534,70	96	99862	42	0	0	0
15	572,09	96	99862	42	0	0	0
17	620,99	99999	1	0	0	0	0