

Aufgabenblatt 1

Physische Speicherstrukturen

- Abgabetermin: **Sonntag, 22.11.2020 (23:59 Uhr)**
- Zur Prüfungszulassung muss ein Aufgabenblatt mit mind. 25% der Punkte bewertet werden und alle weiteren Aufgabenblätter mit mindestens 50% der Punkte.
- Die Aufgaben sollen in Zweiergruppen bearbeitet werden.
- Abgabesystem unter <http://www.dcl.hpi.uni-potsdam.de/submit>
 - ausschließlich pdf-Dateien
 - eine Datei pro Aufgabe namens Aufgabe-<aufgabenNr>.pdf
 - *jedes* Blatt beschriftet mit Namen

Aufgabe 1: Speichermenge und Lesezeit

Gegeben sei eine Festplatte mit folgenden Merkmalen:

- Es gibt fünf Platten mit 9cm Durchmesser, die jeweils zwei Plattenoberflächen besitzen.
 - Jede Plattenoberfläche trägt 20.000 Spuren.
 - Jede Spur besteht durchschnittlich aus 1.000 Sektoren der Größe 512 Byte.
 - 5 % jeder Spur werden von Lücken überdeckt.
 - Die Platten rotieren mit einer Geschwindigkeit von 10.000 U/min.
 - Die Bewegung des Lesekopfes über n Spuren benötigt $(1 + 0,002n)$ ms.
 - Ein Block umfasst 5 Sektoren.
- a) Was ist die Gesamtkapazität der Platte in GB? **1 P**
- b) Wie viele Blöcke liegen auf einem Zylinder? **1 P**
- c) Was ist die durchschnittliche Latenzzeit für einen Block? **2 P**
- d) Auf dieser Festplatte sei eine Datei gespeichert, die eine komplette Spur belegt. Wie lange dauert durchschnittlich das sequentielle Einlesen der gesamten Datei, wenn das Einlesen einer Spur nur an einer bestimmten Position auf der Spur beginnen darf? **2 P**

Aufgabe 2: Two-Phase Multiway Merge Sort

Gegeben sei die Festplatte aus Aufgabe 1 und ein Hauptspeicher von 200 MiB. Wir betrachten in dieser Aufgabe den naiven Two-Phase Multiway Merge Sort Algorithmus, das heißt: In Phase 2 wird jeweils nur der erste Block jeder Teilliste in den Hauptspeicher geladen und jeder Block wird einzeln auf die finale Ausgabeliste geschrieben.

- a) Wie lange dauert das Lesen und Schreiben von Tupeln insgesamt, wenn 20 Millionen Tupel sortiert werden müssen. Ein Block umfasst 100 Tupel. Zur Optimierung der Zugriffsgeschwindigkeit sind die Daten zuvor optimal über die Spuren einer Oberfläche verteilt worden (d.h. die Daten befinden sich sequentiell hintereinander).
HINWEIS: Nutzen Sie die Ergebnisse aus Aufgabe 1! **9 P**
- b) Was ist die Mindestgröße des Hauptspeichers (in Anzahl von Blöcken), um die 20 Millionen Tupel aus Aufgabe a) per Two-Phase Multiway Merge Sort zu sortieren? Begründen Sie Ihre Antwort! **3 P**

Aufgabe 3: Datensätze fester vs. variabler Länge

Gegeben seien die folgenden Anforderungen für eine Relation einer Patientendatenbank: Es gibt 10 Attribute, die immer gefüllt werden (z. B. Name, Geburtsdatum) und weitere 20 optionale Attribute, die nicht für jeden Patienten bekannt oder relevant sind (z. B. aktuelle Blutzuckerwerte). Jedes der optionalen Attribute wird mit der Wahrscheinlichkeit p gefüllt.

Die Werte der erforderlichen Attribute werden in Feldern der definierten Länge von 64 Byte gespeichert, die Werte der optionalen Attribute in Feldern der festen Länge von 8 Byte. Desweiteren nehmen wir an, dass die Tupel lediglich die Daten und keinerlei Header-Informationen speichern.

- a) Was ist die erwartete Größe eines Tupels (in Abhängigkeit von p) unter Verwendung von
- Datensätzen fester Länge? **1 P**
 - Datensätzen variabler Länge? Die optionalen Attribute sollen als getaggte Felder gespeichert werden. Das Tag jedes dieser Attribute umfasst 2 Byte und die Sequenz der Felder wird mit einem gesonderten Tag der Größe 2 Byte abgeschlossen. **1 P**
- b) Für welche Wahrscheinlichkeiten p sind Datensätze fester Länge zu bevorzugen? **2 P**

Aufgabe 4: Speichern von Relationen

Gegeben seien die folgenden beiden Relationen Hersteller und Produkte, wobei Produkte per Fremdschlüssel einem Hersteller zugeordnet sind. Alle Tupel werden als Datensätze fester Länge gespeichert.

Die Hersteller-Relation umfasst 10.000 Tupel. Jeder Datensatz ist 500 Byte lang. Die Produkte-Relation umfasst 40.000 Tupel. Jeder Produkte-Datensatz hat die Größe 200 Byte.

Die Datensätze sollen in Blöcken der Größe 4.096 Byte gespeichert werden, wobei jeder Block 64 Byte für Header-Informationen reserviert. Es gibt keine Datensätze, die auf mehrere Blöcke aufgeteilt sind.

- a) Wie viele Blöcke werden benötigt, um beide Relationen zu speichern, wenn jede Relation am Stück (in getrennte Blockgruppen) gespeichert wird? **4 P**
- b) Wie viele Blöcke werden benötigt, um beide Relationen zu speichern, wenn jeder Hersteller und seine Produkte im selben Block gespeichert werden? Nimm an, dass jeder Hersteller 4 Produkte herstellt. **4 P**
- c) Der Workload auf der Datenbank umfasst zwei Typen von Anfragen:
 - 1) Lesen aller Hersteller-Daten
 - 2) Join der Relationen Hersteller und Produkte

Welche der beiden obigen Speichervarianten (Aufgabenteil a und b) sollten für welche Anfrageart bevorzugt werden? Und warum? (Ein Satz als Antwort genügt.) **4 P**

Aufgabe 5: Verschiedene Lesegeschwindigkeiten

Gegeben sei eine Festplatte, bei der der Radius der äußersten Spur doppelt so groß ist wie der Radius der innersten Spur: $r_A = 2 \cdot r_I$. Wir nehmen an, dass die Größe eines Sektors konstant ist, d.h. ein Sektor auf einer äußeren Spur ist genauso lang und fasst genauso viel Speicher wie ein Sektor auf einer inneren Spur, das gleiche gilt für die Länge der Lücken. Wie viel kleiner ist die Transferzeit eines Sektors auf der äußersten Spur im Verhältnis zur Transferzeit auf der inneren Spur? **2 P**