**מאיצים חישוביים ומערכות מואצות**

**046278**

**דו"ח הגשה – תרגיל בית רטוב 2**

**מתן צחורי 208936989**

**עדי צחורי 315374066**

**חלק 1**

1. ממומש ב-
2. בהרצת השרת במצב Streams עם קיבלנו:  
   נסמנו כ-.
3. נריץ את השרת כעת עם משתנה בקפיצות קבועות בטווח:  
   , נעגל למספר שלם לצורך נוחות.  
     
   להלן התוצאות המתקבלות מההרצות השונות:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Latency** | | | | | **Throughput** | **Load** |
| Maximum | 99th Percent | Median | Minimum | Average |
| 124.3338 | 4.3726 | 0.1737 | 0.0599 | 0.5406 | 43.8 | **43** |
| 123.6057 | 3.1474 | 0.1368 | 0.1241 | 0.2759 | 135.9 | **136** |
| 130.0971 | 2.4609 | 0.1540 | 0.0597 | 0.4467 | 228.7 | **229** |
| 4.9791 | 2.4350 | 0.1415 | 0.0532 | 0.3471 | 320.0 | **321** |
| 4.9230 | 2.8770 | 0.2989 | 0.0428 | 0.5925 | 416.9 | **414** |
| 5.1120 | 2.4287 | 0.7601 | 0.0392 | 0.8746 | 510.6 | **506** |
| 4.9685 | 2.3254 | 0.9517 | 0.0463 | 0.9785 | 599.9 | **599** |
| 4.9347 | 2.5169 | 1.1352 | 0.0483 | 1.1501 | 690.9 | **692** |
| 5.0443 | 2.7071 | 1.1485 | 0.0444 | 1.1559 | 779.2 | **784** |
| 4.8677 | 2.3212 | 1.1628 | 0.0390 | 1.1338 | 870.0 | **877** |

1. להלן מצורף גרף המציג את ההשהיה החציונית עבור קצבי התמונות לשנייה כפי שהגדרנו בסעיף קודם. כאשר ציר ה- הוא ה- וציר ה- הוא ה-:  
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
     
   מגרף זה ניתן לראות כי עבור עומס הקטן מהעומס שמצאנו בסעיף ב, ההשהיה לכל בקשת תמונה נמשכת כמות זמן קטנה יחסית ובפרט נמוכה מההשהיה עבור .  
   אך ככל שהעומס על השרת עולה כך ההשהיה הולכת וגדלה. כלומר ככל שמעמיסים יותר על השרת, כך קצב מילוי הבקשות קטן.

**חלק 2**

1. נחשב את מספר ה- שניתן להריץ עבור כל ונכפיל במספר ה- שקיימים ב- שלנו.  
   את המידע אודות מאפייני ה- נשלוף בעזרת המתודה . המידע אותו נרצה לשלוף הוא: מספר הרגיסטרים הכולל עבור כל , נסמן , הזכרון המשותף הכולל עבור כל , נסמן , ומספר החוטים הכולל עבור כל , נסמן . נרצה גם לדעת מה מספר ה- הכולל שיש לנו ב-, נסמן .  
   נרצה לבדוק איך כל אחד מהמאפיינים הנ"ל מגביל את מספר ה- שניתן להריץ.  
   עבור רגיסטרים נחשב: , כאשר מספר הרגיסטרים לכל חוט נתון והוא 32, ומספר החוטים לכל בלוק גם כן ידוע.  
   עבור זיכרון משותף נחשב: , כאשר את הזכרון המשותף לכל בלוק נמצא באמצעות הדגל (nvcc --ptxas-options=-v) בנוסף לזיכרון המשותף שמשתמשת הפונקציה .  
   עבור חוטים נרצה לבדוק שמספר החוטים המבוקש לא חורג ממספר החוטים האפשריים לבלוק (אם כי אף אחת מהבקשות למספר חוטים בבלוק בתרגיל זה לא חורגת ממספר החוטים האפשרי - 1024). נחשב: .  
   מתוך החישובים הנ"ל ניקח את הערך המינימלי שהוא המגבלה עבורנו ונכפיל במספר   
   ה- ב-. ובסה"כ נקבל את המספר הכולל שנוכל להריץ.
2. ממומש ב-
3. התורים המחברים בין ה- ל- ממומשים עם שנלמד בתרגול.  
   מכיוון בכל צד כותב לצד אחד ורק קורא מהצד השני, פעולות ההכנסה וההוצאה מהתור אינם נדרשים להיות אטומיים.  
   בצד של ה-, ה- הוא היחיד שמכניס ומוציא משני התורים ולכן נוכל לבצע את הפעולות ללא הגנה.  
   בצד של ה-, מספר חוטים שונים יכולים לנסות להוציא/להכניס לתורים באותו זמן, לכן פעולות אלא הן פעולות קריטיות עליהן נדרש להגן עליהן בעזרת מנעול בסקופ ה-. המנעול ממומש בעזרת ו- כפי שראינו בתרגול.
4. עבור 1024 חוטים קיבלנו:
   1. בהרצת השרת במצב עם קיבלנו:  
      נסמנו כ-.
   2. נריץ את השרת כעת עם משתנה בקפיצות קבועות בטווח:  
      , נעגל למספר שלם לצורך נוחות.

להלן התוצאות המתקבלות מההרצות השונות:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Latency** | | | | | **Throughput** | **Load** |
| Maximum | 99th Percent | Median | Minimum | Average |
| 22.1500 | 13.5978 | 0.6356 | 0.0415 | 1.2221 | 2360 | **2356** |
| 32.8370 | 18.8720 | 0.8974 | 0.0427 | 1.6421 | 7322.4 | **7331** |
| 7.4612 | 3.3889 | 1.0118 | 0.0454 | 1.1807 | 12267.9 | **12306** |
| 8.1081 | 3.5187 | 1.1686 | 0.0432 | 1.2836 | 17214.7 | **17280** |
| 25.0771 | 18.4067 | 1.7163 | 0.0413 | 3.2151 | 22136.1 | **22255** |
| 11.8838 | 4.5534 | 1.4339 | 0.0414 | 1.4989 | 27008.2 | **27230** |
| 21.6213 | 15.1227 | 6.7705 | 0.0441 | 7.2534 | 30541.9 | **32205** |
| 129.1646 | 124.1765 | 65.6774 | 0.0495 | 64.7553 | 25235.5 | **37180** |
| 89.8744 | 83.2830 | 42.5507 | 0.0438 | 42.6014 | 30887.6 | **42154** |
| 150.9760 | 144.8387 | 91.2204 | 1.0236 | 80.0443 | 27993.3 | **47129** |

1. עבור 512 חוטים קיבלנו:
   1. בהרצת השרת במצב עם קיבלנו:  
      נסמנו כ-.
   2. נריץ את השרת כעת עם משתנה בקפיצות קבועות בטווח:  
      , נעגל למספר שלם לצורך נוחות.

להלן התוצאות המתקבלות מההרצות השונות:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Latency** | | | | | **Throughput** | **Load** |
| Maximum | 99th Percent | Median | Minimum | Average |
| 22.1209 | 10.0307 | 0.8249 | 0.0469 | 1.2755 | 2583 | **2583** |
| 6.1595 | 3.5298 | 1.0061 | 0.0490 | 1.2279 | 8020.6 | **8037** |
| 32.6071 | 18.9037 | 1.3726 | 0.0471 | 2.2044 | 13359.4 | **13491** |
| 35.8692 | 18.8057 | 1.6893 | 0.0477 | 2.7532 | 18723.4 | **18945** |
| 10.5460 | 5.4561 | 1.5935 | 0.0481 | 1.7014 | 24262.2 | **24399** |
| 34.7279 | 23.4176 | 7.1907 | 0.0517 | 8.3685 | 28093 | **29853** |
| 46.3057 | 37.4587 | 20.0585 | 0.0788 | 19.8775 | 31080.3 | **35307** |
| 182.5065 | 175.2483 | 85.4288 | 0.0583 | 86.0527 | 23637.6 | **40761** |
| 115.9771 | 106.3185 | 53.7820 | 0.3360 | 54.2996 | 30716.9 | **46215** |
| 166.1539 | 159.5514 | 72.1633 | 0.8303 | 76.7883 | 28262.9 | **51669** |

1. עבור 256 חוטים קיבלנו:
   1. בהרצת השרת במצב עם קיבלנו:  
      נסמנו כ-.
   2. נריץ את השרת כעת עם משתנה בקפיצות קבועות בטווח:  
      , נעגל למספר שלם לצורך נוחות.

להלן התוצאות המתקבלות מההרצות השונות:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Latency** | | | | | **Throughput** | **Load** |
| Maximum | 99th Percent | Median | Minimum | Average |
| 28.4818 | 13.0232 | 1.3006 | 0.0695 | 1.7354 | 2986.7 | **2980** |
| 31.0827 | 15.2046 | 1.6722 | 0.0759 | 2.1569 | 9260.9 | **9272** |
| 40.7089 | 22.6278 | 1.9927 | 0.0685 | 3.0714 | 15352.1 | **15565** |
| 40.8086 | 25.5577 | 3.3755 | 0.0695 | 6.3231 | 21610 | **21857** |
| 37.8714 | 20.1226 | 4.3808 | 0.0726 | 5.9098 | 27654.6 | **28149** |
| 64.2646 | 48.5357 | 26.7479 | 0.0817 | 25.5279 | 29493.4 | **34442** |
| 121.6248 | 115.4722 | 63.1935 | 0.1009 | 61.0745 | 27673.6 | **40734** |
| 186.8136 | 171.5398 | 89.3396 | 0.1009 | 88.7153 | 25715 | **47026** |
| 219.0242 | 141.9914 | 72.2073 | 0.6006 | 72.0962 | 24687.8 | **53318** |
| 224.0965 | 214.0395 | 109.9643 | 1.0582 | 109.6600 | 25915.2 | **59611** |

1. להלן מצורפים הגרפים המציגים את ההשהיה החציונית עבור קצבי התמונות לשנייה כפי שהגדרנו בסעיפים קודמים. כאשר ציר ה- הוא ה- וציר ה- הוא ה-: (שני גרפים שונים בגלל סקלות שונות).
2. ניתן ללמוד מתוך הגרפים שכאשר עוברים את נק' ה- עבור כל אחד מתצורות ההפעלה זמני ההשהיה קופצים בצורה דרמתית, והמערכת לא מסוגלת לעמוד בדרישת העומס במצבים אלו.  
   בנוסף, אם נתמקד באזורים בגרף שבהם המערכת כן עומדת בדרישת העומס, ניתן לראות שעבור מספר רק יותר של חוטים ההשהיה קטנה יותר בממוצע.
3. החלפת התורים מזוג תורים לכל המערכת, לזוג תורים לכל יכולה לשפר את ביצועי השרת בהנחה שה- מספיק למלא/לרוקן את התורים בזמן וכן מסוגל לפזר את הבקשות בצורה אחידה בין ה-. כעת, מכיוון שמספר עלולים לגשת לתור להוצאה/הכנסה של בקשות בו זמנית, הגישה לתור מוגנת על ידי מנעול. הפרדת התור לתורים פרטיים יאפשר גישה ללא הגנה עם מנעול (בגלל אופן מימוש התור), ועשוי בכך להאיץ את קצב הטיפול בבקשות – אין חסימת גישה.
4. העברת התור לזכרון ה- עשויה לשפר ביצועים בכך שיהיה חסכון בקריאות מהתור מתוך ה- לזיכרון שנמצא ב-.  
   הגרעין שרץ על כל מנסה בלולאה אינסופית לבדוק אם קיים איבר בתור , ואם כן לשלוף אותו. כלומר מתבצעות מספר גדול מאוד של בקשות גישה לזיכרון הנמצא ב- ובכך מועברות הרבה בקשות זיכרון שעוברות ב- ולוקחות זמן, לעומת זאת ה- ניגש לתור זה בכל פעם שהוא מוסיף בקשה.  
   בהעברת התור לזיכרון ה- רוב הבקשות יגיעו מתוך ה- ולא יצטרכו להמשך מתוך זיכרון ה-.
5. על מנת להעביר את התור לזיכרון של ה-, נצטרך לתת ל- גישה לזיכרון זה.  
   ניתן לבצע זאת באופן סימטרי לאופן שבו אפשרנו ל- לגשת לזכרון של ה-. בעזרת נבקש למפות מרחב כתובות זיכרון ב- לכתובות ב- (מרחב הכתובות יהיה מחוץ לטווח הכתובות הפיזיות של ה-). בבקשת גישה לזיכרון שנמצא ב-, הכתובות יעברו תרגום דרך ה- ונקבל את הכתובת המתאימה ב-.  
   ב- למשל ניתן לבצע זאת בעזרת טכנולוגיית .