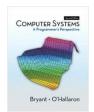
System Level Programming

Unit 10 Performance Measure

李靓 (Li Jing) <u>lijing712@scu.edu.cn</u> School of Computer Science / Software Engineering

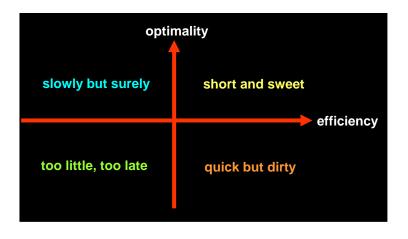




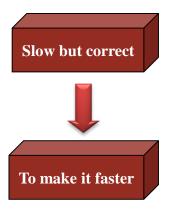
Performance Measure

- ▶ 10.1 Rationale (逻辑依据) for this unit
- ➤ 10.2 Performance Principles (法则)
- 10.3 Performance Measurement
- ▶ 10.4 VC Profiler (评测器)

Attribute of a program

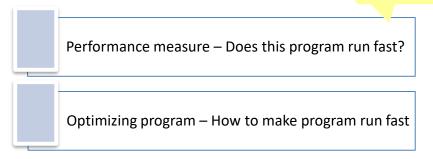


Procedure of developing a program



Some issues need to be considered

This unit

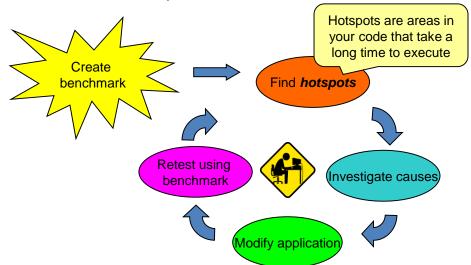


Does this program run fast?

Bottlenecks / Hot spots

We must find what should be optimized

The Software Optimization Process



Performance Measure

- ▶ 10.1 Rationale (逻辑依据) for this unit
- ➤ 10.2 Performance Principles (法则)
- 10.3 Performance Measurement
- ▶ 10.4 VC Profiler (评测器)

> 80/20 rule (Pareto Principle)



In the 1800's an Italian economist, Vilfredo Pareto, identified that 80% of the wealth of the country was owned by 20% of the people.

> 80/20 rule seems to apply to lots of things in life

20% of your stored phone numbers – you call 80% of the time You see 20% of your family and friends – 80% of the time 80% of your monthly pay check vanishes in 20% of the month!!







- > 80/20 rule
 - 80% of CPU time is spent in 20% of the program



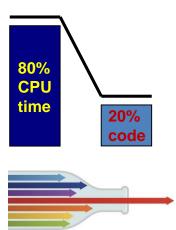
> 80/20 rule

You can have better performance by focusing

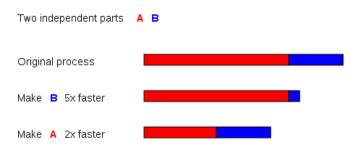
on this 20%

You focus 80% of your time on the vital 20% of things that really matter

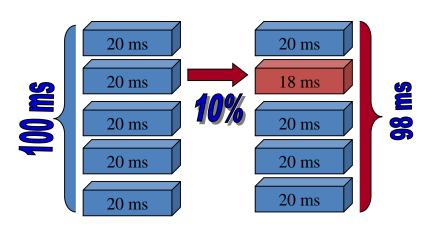
定性 Qualitative



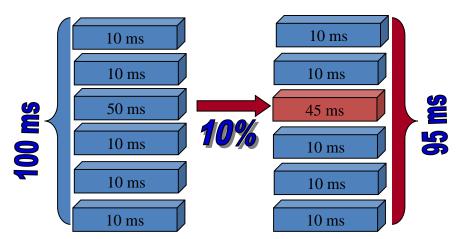
- ➤ Amdahl's Law / Argument (阿姆达尔法则)
 - Named after <u>computer architect</u> <u>Gene Amdahl</u>
 - Used to find the maximum expected improvement to an overall system when only part of the system is improved



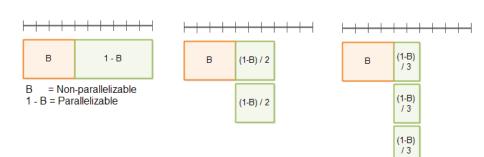
10% on one module means 2% as a whole



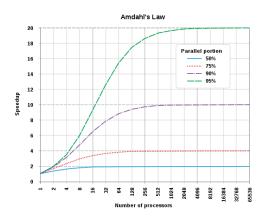
- 10% on one module means 5% as a whole
- Conclusion: focus on module with more CPU time



- ➤ Amdahl's Law / Argument (阿姆达尔法则)
 - Used in <u>parallel computing</u> to predict the theoretical maximum <u>speedup</u>



- ➤ Amdahl's Law / Argument (阿姆达尔法则)
 - Used in <u>parallel computing</u> to predict the theoretical maximum <u>speedup</u>



Suppose the enhancement E accelerates a fraction P of one task by a factor S and the remainder of the task unaffected

$$ExcuteTime(Without E) = 1$$
$$ExcuteTime(With E) = (1 - P) + \frac{P}{S}$$

$$Speedup(E) = \frac{ExcuteTime(Without E)}{ExcuteTime(With E)} = \frac{1}{\left\{ (1 - P) + \frac{P}{S} \right\}}$$

- Quiz
 - We are considering an enhancement to the processor of a web server.
 - The new CPU is 20 times faster on search queries than the old processor.
 - The old processor is busy with search queries 70% of the time.
 - What is the speedup gained by integrating the enhanced CPU?

Performance Measure

- ▶ 10.1 Rationale (逻辑依据) for this unit
- ➤ 10.2 Performance Principles (法则)
- > 10.3 Performance Measurement
- ▶ 10.4 VC Profiler (评测器)

10.3 Performance measurement

- > 10.3.1 What to measure
- 10.3.2 Timing mechanisms
- ➤ 10.3.3 Statistical sampling (统计抽样) / Profiling (分析)

10.3.1 What to measure

> The most common thing to measure is time



10.3 Performance measurement

- > 10.3.1 What to measure
- 10.3.2 Timing mechanisms
- ➤ 10.3.3 Statistical sampling (统计抽样) / Profiling (分析)

10.3.2 Timing mechanisms

- > 10.3.2.1 Introduction
- ➤ 10.3.2.2 Timer in hardware
- > 10.3.3 Timer in OS
- > 10.3.4 Timer in C/C++

➤ What is time? ——1D quantity



"Do you love life? Then don't waste time, because life is made of time."—— Benjamin Franklin

Time is used to

Measure events sequence

Quantify the durations of events and the intervals

上午	第01节课	08:15-09:00
	第02节课	09:10-09:55
	第03节课	10:15-11:00
	第04节课	11:10-11:55
	7701 p 194	11110 111100
下午	第05节课	13:50-14:35
	第06节课	14.45.15.00
		14:45-15:30
	第07节课	15:40-16:25
	第08节课	16:45-17:30
	第09节课	17:40-18:25

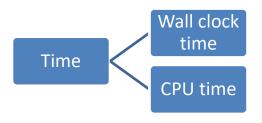
Why we need time in CS?

Hardware need time and timer

OS need time and timer

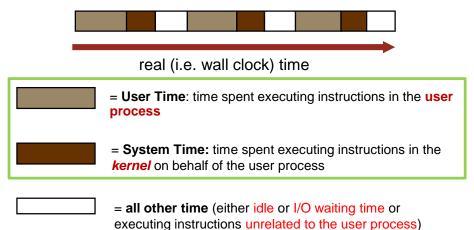
Measuring program performance, e.g. execution time

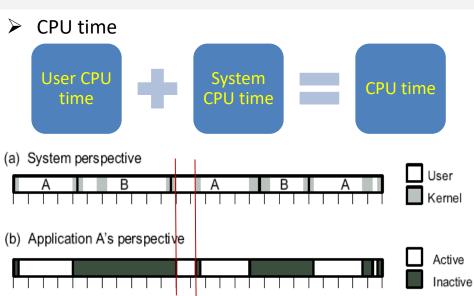
How fast programs run on machine?



The time an ordinary clock on the wall or a wrist watch shows

Wall clock time - The overall time needed to run a particular program or solve a problem.





How to use time in CS?

By using many kinds of timer (定时器)

What is timer

A component in Computer System/CS, as a hardware or software, which can provide the ability of measuring time in some degree.

Different kinds of timer



Timer in hardware
Intel Architecture, 32/X86, 8253



Timer in OS

Windows, GetTickCount() | Linux, jiffies

```
Timer in C/C++
<time.h>
```

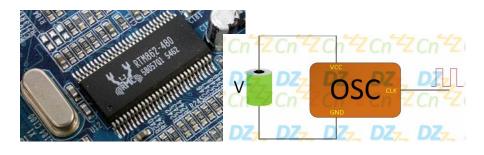
10.3.2 Timing mechanisms

- > 10.3.2.1 Introduction
- ➤ 10.3.2.2 Timer in hardware
- > 10.3.3 Timer in OS
- > 10.3.4 Timer in C/C++

- Real Time CMOS Clock 实时时钟 RTC
 - CMOS RAM: store time & configurations
 - Low powered battery
 - For the boot-up process



- Real Time CMOS Clock 实时时钟 RTC
 - The RTC keeps updating time in the background
 - Source: crystal oscillator produces the original clock frequency



Main frequency | Frequency multiplication

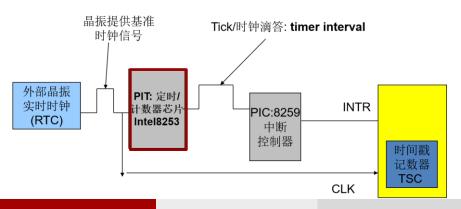
时钟周期/振荡周期 Clock Cycle = seconds per cycle

$$Clock Frequency = \frac{1}{Clock Cycle}$$

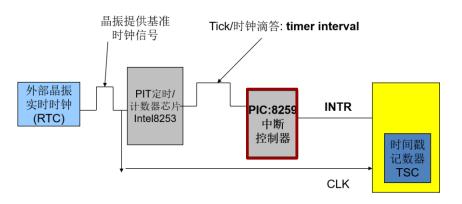
时钟频率 (Hz.= cycle/sec) Clock Frequency = cycles per second

i.e. 1.4GHz

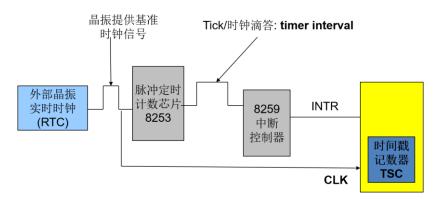
- ▶ PIT (Programmable Interval Timer 可编程间隔定时器)
 - A counter that generates output signal and (may) then trigger an interrupt



PIC (Programmable Interrupt Controller)



- ➤ TSC (Time Stamp Counter 时间戳计数器)
 - TSC receives the CLK signal from RTC
 - A 64-bit register, count the number of cycles



- TSC: Time Stamp Counter
 - Intel Pentium processors (among others) have a very high-speed internal 64-bit counter
 - Be accessed using RDTSC (in nanoseconds)
 - Reads TSC value from registers
- Windows operating systems have an interface to access high-resolution performance counter (in microseconds)

QueryPerformanceFrequency QueryPerformanceCounter

- High-precision timer function Win interface
 - The QueryPerformanceFrequency() function retrieves the frequency of the high-resolution performance counter

```
BOOL QueryPerformanceFrequency (
// address of current frequency
LARGE_INTEGER* lpFrequency );
```

- High-resolution timer function Win interface
 - The QueryPerformanceCounter() function retrieves the current value of the highresolution performance counter

```
BOOL QueryPerformanceCounter (
// Pointer to counter value
LARGE_INTEGER* lpPerformanceCount );
```

LARGE_INTEGER

Be used to represent a 64-bit signed integer value.

For a 64-bit compiler, use **QuadPart** to store 64-bit integer.

Otherwise, use **LowPart** and **HighPart** to store the 64-bit integer.

LARGE_INTEGER example

```
LARGE INTEGER litmp:
LONGLONG QPart1, QPart2:
double dfFreq;
double dfMinus, dfTim;
QueryPerformanceFrequency(&litmp):
dfFreg = (double)litmp.QuadPart; // 获得计数器的时钟频率
QueryPerformanceCounter(&litmp):
QPart1 = litmp.QuadPart; // 获得初始值
Sleep(100);
QueryPerformanceCounter(&litmp):
QPart2 = litmp.QuadPart; //获得中止值
dfMinus = (double)(QPart2-QPart1);
dfTim = dfMinus / dfFreq: // 获得对应的时间值,单位为秒
```

- Example code
 - Using the high precision timer under Windows can be found in the .zip file precise.zip

```
void precise_start()

Begin timing

double precise_stop()

To get the elapsed time in seconds
```

10.3.2 Timing mechanisms

- > 10.3.2.1 Introduction
- ➤ 10.3.2.2 Timer in hardware
- > 10.3.3 Timer in OS
- > 10.3.4 Timer in C/C++

Windows System Time

- The number of milliseconds elapsed since the system was last started
- DWORD GetTickCount(void); //Cycles every 49.7 days
- Read from RTC and then convert

SYSTEMTIME

- GetLocalTime(), 1ms
- GetSystemTime() UTC, Universal Time Coordinated

FILETIME

- GetSystemTimeAsFileTime(), $100ns = 0.1\mu m$
- Win 1601.1.1 | Linux 1970.1.1

```
typedef struct _SYSTEMTIME {
   WORD wYear; // The valid values are 1601 through 30827.
   WORD wMonth;
   WORD wDayOfWeek;
   WORD wDay;
   WORD wHour:
   WORD wMinute:
   WORD wSecond;
   WORD wMilliseconds;
SYSTEMTIME, *PSYSTEMTIME;
typedef struct FILETIME {
   DWORD dwLowDateTime;
   DWORD dwHighDateTime;
} FILETIME. *PFILETIME:
```

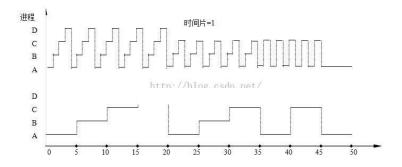
```
#include <windows.h>
#include <stdio.h>
void main() {
  SYSTEMTIME st, lt;
  GetSystemTime(&st);
  GetLocalTime(&lt);
  printf("The system time is: %02d:%02d\n", st.wHour, st.wMinute);
  printf(" The local time is: %02d:%02d\n", lt.wHour, lt.wMinute);
```

// Sample output
The system time is: 14:34
The local time is: 22:34

```
int main()
   ULARGE INTEGER uli;
    FILETIME
                  ft;
    SYSTEMTIME
                    st;
   GetSystemTimeAsFileTime(&ft);
    uli.LowPart = ft.dwLowDateTime;
    uli.HighPart = ft.dwHighDateTime;
    printf("System File Time: %I64u\n", uli.QuadPart);
    FileTimeToSystemTime(&ft, &st);
    printf("System Time (YYYY-MM-DD HH:MM:SS): %d-%d-%d %d:%d:%d\n",
        st.wYear, st.wMonth, st.wDay, st.wHour, st.wMinute, st.wSecond);
    return EXIT SUCCESS:
System File Time: 132176695123942571
System Time (YYYY-MM-DD HH:MM:SS): 2019-11-8 6:51:52
Press any key to continue
```

Time slice

A specific number of clock ticks (时钟单元) before process gets moved to another state



10.3.2 Timing mechanisms

- ➤ 10.3.2.1 Introduction
- ➤ 10.3.2.2 Timer in hardware
- > 10.3.3 Timer in OS
- > 10.3.4 Timer in C/C++

clock_t clock()

```
#include <time.h>
        void elapsed time()
           printf( "%d ms\n", clock());
           printf( "%d s\n", clock()/CLOCKS_PER_SEC);
How long the calling process has spent.
clock() The processor time that have elapsed
1 clock tick = 1/CLOCKS PER SEC
```

- ➤ Macro宏 CLOCKS_PER_SEC
- Data types clock_t

```
// time.h
#define CLOCK PER SECOND((clock t)1000)
#ifndef CLOCK T DEFINED
typedef long clock t;
#define CLOCK T DEFINED
#endif
```

clock() in C/C++

```
Time for 1000000 iterations: 0.069s
Precision is 1000 clocks per second.
Press any key to continue
```

```
int main(int argc, char* argv[]) {
     long n = 1000000;
    clock t start = clock();
    my subroutine(n);
    clock t finish = clock();
```

return 0;

```
#include <stdlib.h>
                                 #include <time.h>
                                 #include <iostream.h>
                                 void my subroutine(long n) {
                                       // timing a subroutine call:
                                       char s[16]:
                                       for (long i = 0; i < n; i++) {
                                            _itoa(i, s, sizeof(s));
double duration = (double)(finish - start) / CLOCKS PER SEC;
cout << "Time for " << n << " iterations: " << duration << "s"
          <<endl<< "Precision is " << CLOCKS PER SEC
          << " clocks per second." << endl;
```

time_t time() Calendar time

```
int main(void)
  time tlt;
  It = time(NULL);
  printf("The Calendar Time now is %d\n",lt);
  return 0;
 The Calendar Time now is: 1573198707
 Press any key to continue
从1970年1月1日0时0分0秒到现在的秒数
```

Data types	time_t	struct tm
Function	time()	localtime() gmtime()

```
// time.h
#ifndef _TIME_T_DEFINED
typedef long time_t;
#define _TIME_T_DEFINED
#endif
```

structure tm

```
#ifndef TM DEFINED
struct tm {
    int tm_sec;
    int tm_min;
    int tm_hour;
    int tm_mday;
    int tm mon;
    int tm_year;
    int tm_wday;
    int tm_yday;
    int tm isdst;
};
#define _ TM_DEFINED
#endif
```

structure tm

Fri Nov 08 16:00:35 2019 Fri Nov 08 16:00:35 2019 Fri Nov 08 08:00:35 2019 Press any key to continue

```
int main(void)
  time t lt:
  It = time(NULL); // calendar time
  printf("%s\n", asctime(&lt)); // convert lt to string
  struct tm *ptr;
  ptr = localtime(&lt); // convert calendar time to local tm
  printf("%s\n", asctime(ptr)); // convert ptr to string
  ptr = gmtime(&lt); // convert calendar time to greenwich tm
  printf("%s\n", asctime(ptr)); // convert ptr to string
  return 0:
```

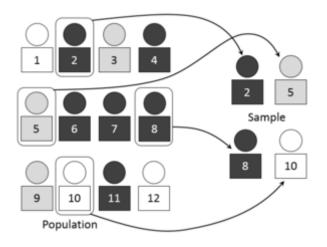
> Functions

	Functions	Description
clock_t	clock()	Process time
time_t	time() difftime() gmtime() localtime()	Get time Time difference To GMT time To local time
struct tm	mktime() strftime()	To time_t Format

10.3 Performance measurement

- 10.3.1 What to measure
- 10.3.2 Timing mechanisms
- ➤ 10.3.3 Statistical sampling (统计抽样) / Profiling (分析)

➤ Statistical Sampling 统计抽样法



- ➤ Statistical Sampling 统计抽样法
 - Any sampling procedure that uses the laws of probability to measure status of the program.



A timer periodically interrupts the program and records the program counter Estimate where time is spent in the program

Check if the program spends most of its time in a few places

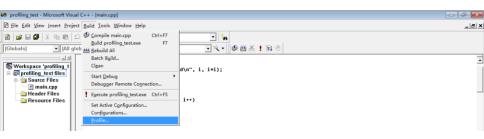
> Why

- It saves time and involves less cost
- In some cases, it might not be possible to check 100% (risk)



- > Software profiler
 - A program that benchmarks the execution of one or more pieces of procedural code to help the user understand where the time is being spent in terms of code execution.
 - GNU Gprof Linux
 - Vtune Intel
 - Visual C++ profiler
 - Valgrind for different platforms

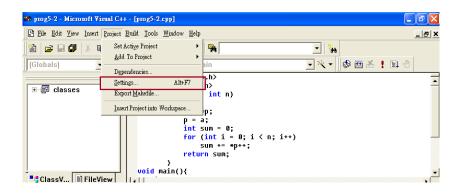
- VC profiler
 - In Visual C++, the Profile... entry in the Build menu gives instructions on obtaining a profile.
 - Profiling is only available in the Professional and Enterprise editions of Visual C++.



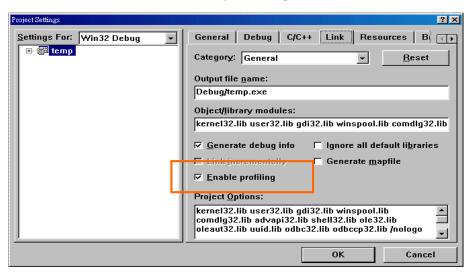
Performance Measure

- ▶ 10.1 Rationale (逻辑依据) for this unit
- ➤ 10.2 Performance Principles (法则)
- > 10.3 Performance Measurement
- ▶ 10.4 VC Profiler (评测器)

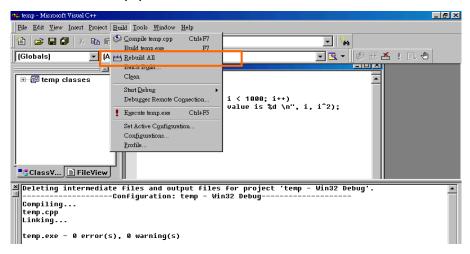
Procedure (1) – setting



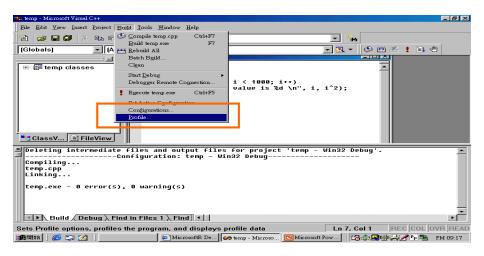
Procedure (2) – enable profiling



Procedure (3) – rebuild



Procedure (4) – run with profiling



```
#include <stdio.h>
#include <stdlib.h>
void main()
  int i;
  for (i = 0; i < 1000; i++)
     printf("The value is %d\t %d \n", i, i*i);
```

```
#include <stdio.h>
#include <stdib.h>
void main()
{
int i;
for (i = 0; i < 1000; i++)
printf("The value is %d\t %d \n", i i*i);
}
```

Result of a simple for loop — total time is 248 ms

```
Time in module: 248.623 millisecond
Percent of time in module: 180.0%
Functions in module: 1
Hits in module: 1
Module function coverage: 180.0%

Func Func+Child Hit
Time % Time % Count Function

248.623 180.8 248.623 180.8 1 _main (test.obj)
```

```
#include <stdio.h>
#include <stdib.h>
void main()
{
   int i;
   for (i = 0; i < 1000; i++)
        printf("The value is %d\t %d \n", i i*i*i);
}

**ClassView | FileView | **Include <stdio.h>
#include <stdio.h
#includ
```

Result of a simple for loop — total time is 268 ms

```
Module Statistics for test.exe

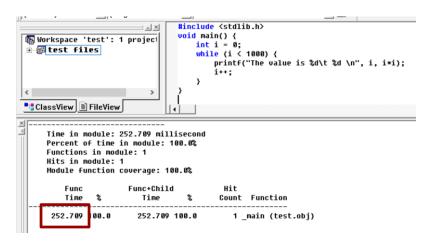
Time in module: 268.222 millisecond
Percent of time in module: 100.0%
Functions in module: 1
Hits in module: 1
Hodule function coverage: 100.0%

Func Func+Child Hit
Time % Time % Count Function

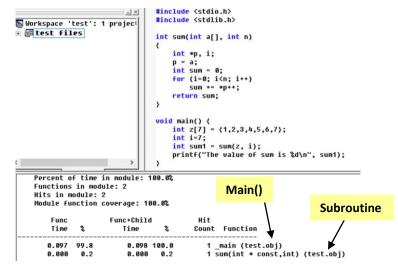
268.222 100.0 268.222 100.0 1 _main (test.obj)
```

```
#include <stdio.h>
#include <stdlib.h>
void main() {
       int i = 0;
       while (i < 1000) {
              printf("The value is %d\t %d \n", i, i*i);
              i++;
```

Example – result in millisecond second



Example with a sub-routine



 A program that can be used to determine million floating point operations (MFLOPS)

```
// This is matrix multiplication
#include <stdio.h>
#include <stdlib.h>
void main(){
          float a[250][250], b[250][250], c[250][250];
          int i, j, k;
          for (i = 0; i < 250; i++)
               for (j = 0; j < 250; j++)
                   for (k = 0; k < 250; k++)
                        // matrix multiplication
                        c[i][j] += a[i][k] * b[k][j];
```

```
Program Statistics
    Command line at 2019 Nov 08 16:43: "D:\work\SLP\2019\week11\test\Debug\test"
    Total time: 110.217 millisecond
    Time outside of functions: 3.500 millisecond
    Call depth: 1
    Total functions: 1
    Total hits: 1
    Function coverage: 100.0%
    Overhead Calculated 1
    Overhead Average 1
Module Statistics for test.exe
    Time in module: 106.708 millisecond
    Percent of time in module: 100.0%
    Functions in module: 1
    Hits in module: 1
    Module function coverage: 180.8%
        Func
                     Func+Child Hit
Time % Count Function
        Time %
     106.708 100.0 106.708 100.0 1 _main (test.obj)
```

Determination of Mega Flop

```
for (i = 0; i< 250; i++)

for (j = 0; j < 250; j++)

for (k =0; k <250; k++)

// matrix multiplication

c[i][j] += a[i][k] * b[k][j];
```

- 250 x 250 x 250 = 15625000 floating point operations
- The performance: 15625000/106ms = 15.625 x 10^6 /0.106 s = 147 MFLOPs (mega floating point operation).
- Try your computer

Same output but change the program

```
// this program uses a temporary location t to store the value
void main() {
   float a[250][250], b[250][250], c[250][250];
   int i, j, k;
   float r = 0.0;
   for (i = 0; i < 250; i++) {
       for (j = 0; j < 250; j++) {
          for (k = 0; k < 250; k++)
              r += a[i][k] * b[k][j]; //this is matrix multiplication
             c[i][i] = r;
```

Same machine – 90ms, why?

This is related to the cache memory effect, as the data is stored in cache.

```
Hodule Statistics for test.exe

Time in module: 90.678 millisecond
Percent of time in module: 100.0%
Functions in module: 1
Hits in module: 1
Hodule function coverage: 100.0%

Func Func+Child Hit
Time % Time % Count Function

90.678 100.0 90.678 100.0 1 _main (test.obj)
```

A profiler is a great way to find where bottlenecks occur, so we can make our code more efficient, It can:

```
Command line at 2019 Nov 08 16:54: "D:\work\SLP\2019\week11\test\Debug\test"
    Total time: 93.231 millisecond
    Time outside of functions: 2.554 millisecond
                                                           Source file: d:\work\slp\2019\week11\test\test.cpp
    Call depth: 1
    Total functions: 1
                                                             Line Covered Source
    Total hits: 1
    Function coverage: 100.0%
                                                                          uoid main() (
    Overhead Calculated 0
                                                                               float a[250][250], b[250][250], c[250][250];
                                                                               int i. j. k:
    Overhead Average 0
                                                                               float r = 0.0;
                                                                               for (i = 0; i< 250; i++) {
Module Statistics for test.exe
                                                                                   for (j = 0; j < 250; j++) {
                                                                                        for (k =0; k <250; k++)
    Time in module: 90.678 millisecond
                                                                                             r += a[i][k] * b[k][i]; //thi
    Percent of time in module: 100.0%
                                                                                            c[i][i] = r;
    Functions in module: 1
    Hits in module: 1
                                                              11:
    Module function coverage: 100.0%
                                                              12:
                       Func+Child
                                             Hit
        Func
        Time
                          Time
                                            Count Function
                                               1 main (test.obj)
      98.678 188.8
                          90.678 100.0
```

- Profiling is used to help programmers to identify
 - Which areas of the program are causing sluggish bottlenecks.
 - ✓ Which parts of the code are being called the most often.