2017.1.11

中文答题；满分 70 分；考试时间两个半小时

1. 名词解释（每个 2 分，共计 16 分）

Hazard	Loop Unrolling	Exception	Reorder Buffer
Data Level Parallelism	栅栏同步	网络直径	等分带宽

2. （5 分）大规模机器的同步有哪些软件和硬件支持方法？

3. （13 分）什么是多处理机的相关性（coherency）和一致性（consistency）？给出解决相关性的监听协议的工作原理。

4. （7 分）In the following loop, find all the true dependences, output dependences, and anti-dependencies. Eliminate the output dependences and anti-dependences by renaming.

```
for (i = 0; i < 100; i++) {
    A[i] = A[i] * B[i];      /*  S1  */
    B[i] = A[i] + c;         /*  S2  */
    A[i]= C[i] * c;          /*  S3  */
    C[i] = D[i] * A[i];      /*  S4  */
}
```

5. （14 分）We begin by looking at a simple two-issue, statically-scheduled superscalar MIPS pipeline, using the pipeline latencies from the table.

Instruction producing result	Instruction using result	Latency in clock cycles
FP ALU op	Another FP ALU op	3
FP ALU op	Store double	2
Load double	FP ALU op	1
Load double	Store double	0

This processor can issue two instructions per clock cycle, where one of the instructions can be a load, store, branch, or integer ALU operation, and the other can be any floating-point operation. For the following code:

```
Loop:   L.D       F0, 0(R1)       ;F0 = array element
        ADD.D     F4, F0, F2      ;add scalar in F2
        S.D       F4, 0(R1)       ;store result
        DADDUI    R1, R1, #-8     ;decrement pointer, 8 bytes(per DW)
        BNE       R1, R2, Loop    ;branch if (R1 != R2)
```

(1) How many clock cycles of the loop per element?
(2) Unroll this loop to make five copies and write the unrolled and scheduled code of the loop in this processor.
(3) Calculate the factor of performance improvement.
(4) Show a software-pipeline version of this loop, which increment stall the elements of

an array whose starting address is in R1 by the contents of F2. You MUST include the start-up and clean-up code.

6. （5 分）Assuming a hypothetical GPU with following characteristics:

- Clock rate 1.5GHz
- Contents 16 SIMD processors, each containing 16 single-precision floating-point units
- Has 100 GB/sec off-chip memory bandwidth

Without considering memory bandwidth, what is the peak single-precision floating-point throughput for this GPU in GFLOP/sec, assuming that all memory latencies can be hidden? Is this throughput sustainable with the given memory bandwidth limitation? Why?

7. （10 分）Consider a branch-target buffer that has penalties of zero, two and two clock cycles for correct conditional branch prediction, incorrect prediction, and a buffer miss, respectively. Consider a branch-target buffer design that distinguishes conditional and unconditional branches, storing the target address for a conditional branch and the target instruction for an unconditional branch.

(1) Assuming a 90% hit rate, 90% accuracy, and 15% branch frequency. How much faster is the processor with the branch-target buffer versus a processor that has a fixed two-cycle branch penalty?

(2) What is the penalty in clock cycles when an unconditional branch is found in the buffer?

(3) Determine the improvement from branch folding for unconditional branches. Assume a 90% hit rate, an unconditional branch frequency of 5%, and a two-cycle penalty for a buffer miss. How much improvement is gained by this enhancement? How high must the hit rate be for this enhancement to provide a performance gain?