Undergraduate Computer Architecture, Fall 2024
Midterm Exam, 2024-11-05

If you agree with the following sentence, please sign your name below it. (If you take this exam remotely, please copy it to your answer sheet and sign on the answer sheet).
I have not cheated nor have I received any help from other students in the exam.

Student ID: b11902038 *Name:* 鄧博允

1.

(a) instruction-level parallelism can be done with pipelining, which is basically execute multiple instruction in parallel (that is, at the same clock cycle).

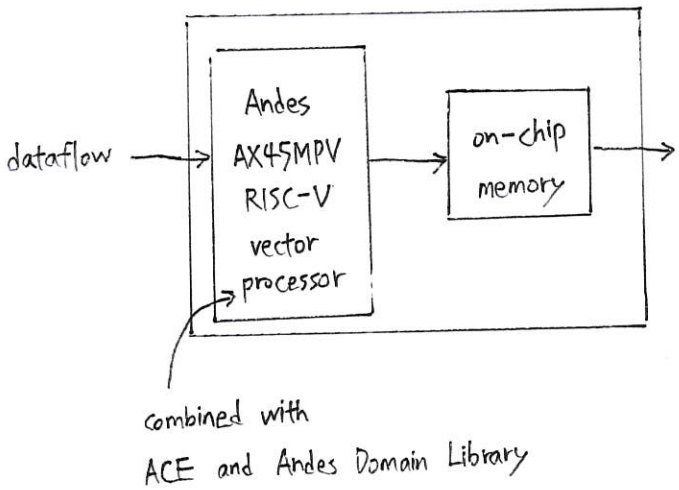
(b) yield is a process

(c) dynamically linked library is used in a technique called "dynamic linking", where the compiler decide which library to use and link them to the program.
dynamically

(d) VLIW(Very Long Instruction Word) is a group of instructions which grouped by the compiler to perform static multiple issue. Generally, this is decided by pipeline resources required.

2.

(e)



(f) this new chip will use vector processor and then execute most of the operation in on-chip memory.

(g) higher energy efficiency

improve latency on inference task

faster on running AI models : 100 x faster

cheaper : 10x cheaper

(h) Fractile plan to use Andes's vector processing unit on their new AI chips, by integrate it with Fractile's in-memory computing architecture via ACE.

(i) CPI stands for cycles needed per instruction

(k) $\text{SPECratio} = \dfrac{\text{Execution time}}{\text{Execution} + \text{Reference time}}$

3.

(i, k)

Description	Name	Instruction Count x 10^9	CPI	Clock cycle time (seconds x 10^{-9})	Execution Time (seconds)	Reference Time (seconds)	SPECratio
Perl interpreter	perlbench	2684	0.42	0.556	627	1774	0.26
GNU C compiler	gcc	2322	0.67	0.556	863	3976	0.18
Route planning	mcf	1786	1.22	0.556	1215	4721	0.20
Discrete Event simulation - computer network	omnetpp	1107	0.82	0.556	507	1630	0.24
XML to HTML conversion via XSLT	xalancbmk	1314	0.75	0.556	549	1417	0.28
Video compression	x264	4488	0.33	0.556	813	1763	0.32
Artificial Intelligence: alpha-beta tree search (Chess)	deepsjeng	2216	0.57	0.556	698	1432	0.33
Artificial Intelligence: Monte Carlo tree search (Go)	leela	2236	0.79	0.556	987	1703	0.37
Artificial Intelligence: recursive solution generator (Sudoku)	exchange2	6683	0.46	0.556	1718	2939	0.37
General data compression	xz	8533	1.33	0.556	6290	6182	0.50
mean	–	–	–	–	–	–	

(j)

Benchmarks like mcf and xz have relatively poor performance, this may be attribute to the type of instructions used in the task. For example, when performing general data compression, it may access memory very often.

(l)

$$\text{mean SPECration} = \frac{\text{total execution time}}{\text{total execution + reference time}}$$

(m)

- the instruction count remain the same, since it does not depend on clock frequency
- the clock cycle time can be calculate by $\frac{1}{\text{clock frequency}} = 0.278\ (\text{seconds} \times 10^{-9})$
- the CPI, execution time are also decrease by $\frac{1}{2}$
- the reference time remain the same

(n)

- execution time decrease by $\frac{1}{2}$
- other remain the same

4.

(o) if there's no dependency between operations, we can start four operations since there's four groups of components: one for read/write operation; one for integer arithmetic operation; one for integer multiply/div operation; one for float arithmetic operation

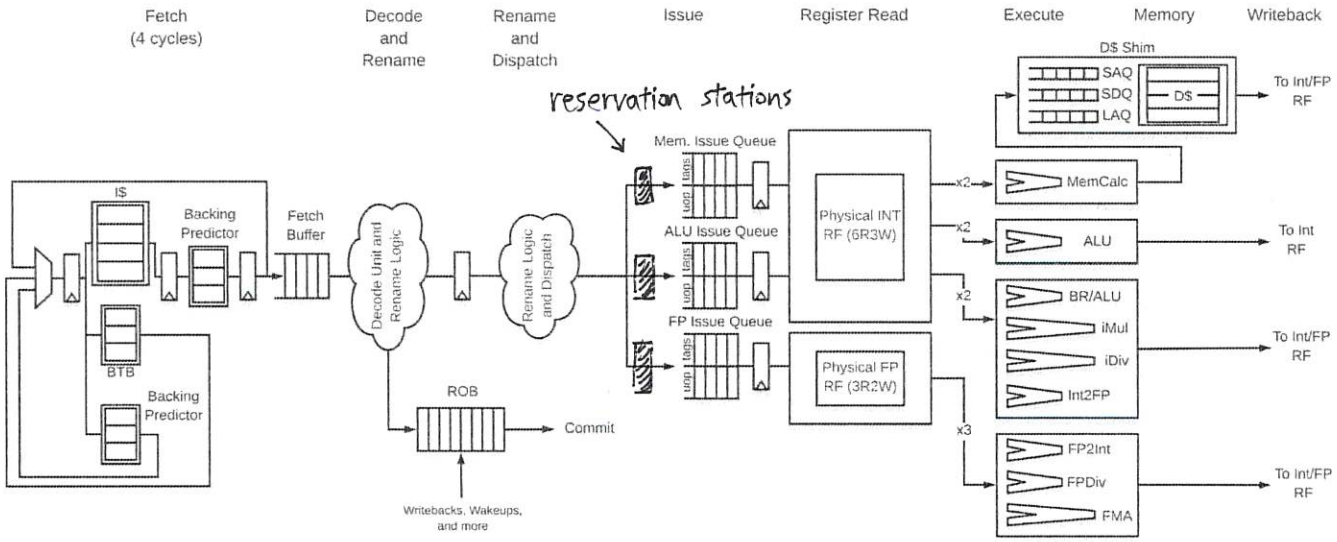
For questions (p) and (q), you may draw on the diagram to support your answer.

(p)

branch prediction occur in Fetch stage, where it make prediction based on BTB. If mispredicted, it will re-fetch the next instruction after calculate address at Execute stage, which will cause bubbles on stage Fetch ~ Execute.

(q)

reservation stations are shown in the diagram, where the pipeline can reserve operands of three queues, making the CPU can execute them out of order.



(r)

Both EX, MEM hazards can occur if some operation is using the result of the operation one step before.

We can reduce the bubbles by forwarding the result of EX, MEM to internal buffers, where the next operation can directly use these result instead of stall to wait it stored to memory.

(s)

When multiple exceptions occur, they are handle by the handler, while at the same time, the pipeline can execute those instructions who are not affected, then store the result at reservation stations instead of commit it. When the exception handler return, the pipeline can check the result correct or not, and commit them if everything alright.

Please write down your student Id and your name here.

Student ID: b11902038 *Name:* 鄧博元