# 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 & Name*:

1. **(20 pts) [Basic Knowledge]** Please explain the following terms as details as possible, with no more than 50 words.

   (a) (5 pts) instruction-level parallelism

   (b) (5 pts) yield (as in chip manufacturing process)

   (c) (5 pts) dynamically linked library

   (d) (5 pts) VLIW

2. **(25 pts) [ISA and RISC-V]** The press release shown on the following 2 pages, titled *Fractile Licenses Andes Technology's RISC-V Vector Processor as It Builds Radical New Chip to Accelerate AI Inference*, describes a recent trend which seems very positive to the RISC-V community. As you have gained certain knowledge about RISC-V during this course, you should be able to comprehend and analyze this article. Please answer the following questions based on the information provided by the report and the knowledge learned from this course.

   (e) (5 pts) In this article, Fractile is building a new chip. Please describe the hardware architecture of this new chip by drawing a picture and include the components mentioned in this article as much as you can.

   (f) (5 pts) Suppose you are building an AI system on the new chip. There is an AI application running on top of a Linux operating system. The AI application receives text input from the user and runs a large language model to perform inference. Please describe how this new chip accelerates the AI application by drawing the software components and data flow on the picture in (e).

   (g) (5 pts) Based on this article, what are the advantages of this new chip? Please mention all the advantages with numbers to show the advantages, e.g. Shorter latency: XX times faster.

   (h) (10 pts) The article mentions that Fractile is betting on inference scaling as the next frontier of AI scaling. Please use your words to summarize how Fractile plans to collaborate with Andes Technology to face for the challenges of this next frontier. Don't just copy and paste. Your summary should remove those advertising/self-promoting words in the press release and describe the trends, challenges, and solutions in more precise technical terms.

(Article from https://www.andestech.com/en/2024/10/22/fractile-licenses-andes-technologys-risc-v-vector-processor-as-it-builds-radical-new-chip-to-accelerate-ai-inference/)

# Fractile Licenses Andes Technology's RISC-V Vector Processor as It Builds Radical New Chip to Accelerate AI Inference

**San Jose, CA — Oct. 22, 2024** — Andes Technology, a leading supplier of high-efficiency, low-power 32/64-bit RISC-V processor cores and Founding Premier member of RISC-V International, are proud to announce a partnership with Fractile, the company building the chips and systems needed to reach the next frontier of AI performance. Fractile is developing AI inference accelerators based on in-memory compute, and aim to be able to run frontier AI models – large language, vision and audio models – two orders of magnitude faster than existing hardware, at a tenfold reduction in cost.

Large language models and other foundation models have become the driving force behind the skyrocketing scale of data center AI compute requirements. From ChatGPT to the open-source Llama model series, LLMs and other foundation models are finding widespread application. Model inference – the process of serving these trained models – is coming to be the dominant portion of compute costs, exceeding the cost of model training. Fractile has licensed the powerful Andes AX45MPV RISC-V vector processor, combined with ACE (Andes Automated Custom Extension™) and Andes Domain Library, and plans to incorporate the vector processing unit into their first generation data center AI inference accelerator.

Fractile's uses novel circuits to execute 99.99% of the operations needed to run model inference in on-chip memory. This removes the need to shuttle model parameters to and from processor chips, instead baking computational operations into memory directly. This architecture drives both much higher energy efficiency (TOPS/W) as well as dramatically improved latency on inference tasks (tokens per second per user in an LLM context, for instance). The company has been betting on inference scaling – leveraging more inference time-compute to improve AI performance – as the next frontier of AI scaling. The AI world seems to agree, with OpenAI recently releasing their latest LLM, o1, which requires orders of magnitude more inference compute than previous LLMs. Fractile's hardware and software stack is built to take models that can still take many seconds to produce an answer on current hardware, and make this instantaneous.

As part of the collaboration, Fractile will integrate Andes Technology's high-performance RISC-V vector processor with its own groundbreaking in-memory computing architecture via ACE. Fractile's architecture leverages the strengths of both companies, aiming to deliver an exceptionally fast and cost-effective AI inference system that overcomes the limitations of conventional computing methods – blasting through the memory bottleneck.

Dr. Charlie Su, President and CTO of Andes Technology, expressed his enthusiasm for the partnership, "AX45MPV, with strong compute capabilities, high memory bandwidth and the flexible ACE tool, has been chosen by innovative AI companies large and small since its debut in 2023. Andes RISC-V vector processors have enabled many AI SoCs to break free from architecture limitation and achieve new levels of performance and efficiency. We are confident that the synergy between Fractile's In-Memory Computing technologies and Andes' award-winning RISC-V vector processing will lead to yet another success."

Dr. Walter Goodwin, CEO and founder of Fractile, added: "The limitations of existing hardware present the biggest barrier to AI performance and adoption. Andes Technology has unmatched technical and commercial leadership on RISC-V vector processors and is a natural partner for us as we build Fractile's accelerator systems. Building hardware for AI acceleration is intrinsically hard – the world's leading models can change overnight, while chips take time to bring to market. Software-programmable vector processors like Andes' are a key part of staying robust to these changes. We're delighted to announce this collaboration as Fractile furthers its mission to supercharge inference.

For more information about Andes Technology and Fractile, please visit their respective websites at www.andestech.com and www.fractile.ai.

**About Andes Technology**

Nineteen years in business and a Founding Premier member of RISC-V International, Andes is a publicly-listed company (TWSE: 6533; SIN: US03420C2089; ISIN: US03420C1099) and a leading supplier of high-performance/low-power 32/64-bit embedded processor IP solutions, and the driving force in taking RISC-V mainstream. Its V5 RISC-V CPU families range from tiny 32-bit cores to advanced 64-bit Out-of-Order processors with DSP, FPU, Vector, Linux, superscalar, and/or multi/many-core capabilities. By the end of 2023, the cumulative volume of Andes-Embedded™ SoCs has surpassed 14 billion. For more information, please visit https://www.andestech.com. Follow Andes on LinkedIn, Twitter, Bilibili and YouTube!!

**About Fractile**

Fractile is an AI hardware company that is building its first groundbreaking new AI chip, capable of running state-of-the-art AI models up to 100x faster and 10x cheaper than existing hardware. Founded in 2022 in London by 28-year-old artificial intelligence PhD Walter Goodwin, Fractile's transformative computing technology will enhance collective AI capabilities by enabling the largest and most capable neural networks of today and tomorrow to run faster, more efficiently and more sustainably. The company has raised $17.5m (£14m) in funding from investors including the NATO Innovation Fund, Kindred Capital, Oxford Science Enterprises, Cocoa and Inovia Capital, as well as angel investors including Hermann Hauser (founder, Acorn, Amadeus), Stan Boland (ex-Acorn, Icera, NVIDIA and Five AI) and Amar Shah (co-founder, Wayve).

14/6

$CPI = \frac{C}{I} = \frac{\text{exec Time} / \text{c time}}{\text{IN ut}} = \frac{2s}{y \cdot 10^9 c} \quad \frac{2}{xy}$

3. **(30 pts) [Computer Performance]** The table below shows the results of SPECspeed 2017 Integer benchmarks running on a computer with a 1.8GHz Intel Xeon E5-2650L. Note that some of the text/numbers in the table are missing. Please answer the following questions:

| | | X | | y | z | MEM | |
|---|---|---|---|---|---|---|---|
| | | Instruction Count (billion) | CPI | Clock Cycle Seconds ($\times 10^{-9}$) | Execution Time (Seconds) | Reference Time (Seconds) | SPECratio |
| Perl interpreter | perlbench | 2684 | 0.42 | 0.556 | 627 | 1774 | |
| GNU C compiler | gcc | 2322 | 0.66 | 0.556 | 863 | 3976 | |
| Route planning | mcf | 1786 | 1.22 | 0.556 | 1215 | 4721 | |
| Discrete Event simulation - computer network | omnetpp | 1107 | 0.82 | 0.556 | 507 | 1630 | |
| XML to HTML conversion via XSLT | xalancbmk | 1314 | 0.75 | 0.556 | 549 | 1417 | |
| Video compression | x264 | 4488 | 0.32 | 0.556 | 813 | 1763 | |
| Artificial Intelligence: alpha-beta tree search (Chess) | deepsjeng | 2216 | 0.56 | 0.556 | 698 | 1432 | |
| Artificial Intelligence: Monte Carlo tree search (Go) | leela | 2236 | 0.79 | 0.556 | 987 | 1703 | |
| Artificial Intelligence: recursive solution generator (Sudoku) | exchange2 | 6683 | 0.46 | 0.556 | 1718 | 2939 | |
| General data compression | xz | 8533 | 1.32 | 0.556 | 6290 | 6182 | |
| mean | – | – | – | – | – | – | |

(i) (5 pts) Please explain what CPI is and fill in the missing CPI fields. Do not use more than 2 digits for each field.

(j) (5 pts) Based on the filled CPI numbers, which benchmarks are executed by the Intel Xeon E5-2650L with relatively ~~poor performance~~ among these benchmarks? Which factors can result in poor CPI?

(k) (5 pts) The SPECratio is intended as an indicator for performance. Please explain how SPECratio is calculated for each benchmark and fill in the missing SPECratio field for each benchmark. Do not use more than 2 digits for each field.

(l) (5 pts) Please explain how the mean SPECratio is calculated and write down the equation to calculate the number.

(m)(5 pts) Suppose the number of pipeline stages of the Intel Xeon E5-2650L CPU is doubled to increase the clock frequency to ~~3.6GHz. How~~ will this change affect the results of the same benchmarks? For each field, discuss how and how much it would be changed.

(n) (5 pts) Suppose the number of pipelines of the Intel Xeon E5-2650L CPU is doubled. How will this change affect the results of the same benchmarks? P For each field, discuss how and how much it would be changed.

4 inst/cycle

4 inst/4 cycle

4. **(25 pts) [Dynamically Scheduled Pipeline]** The figure in the next page shows a RISC-V processor pipeline in an open-source project called the Berkeley Out-of-Order Machine (BOOM). Conceptually, BOOM is broken up into 10 stages: Fetch, Decode, Register Rename, Dispatch, Issue, Register Read, Execute, Memory, Writeback and Commit. However, many of those stages are combined in the current implementation, yielding seven stages: Fetch, Decode/Rename, Rename/Dispatch, Issue /RegisterRead, Execute, Memory, and Writeback (Commit occurs asynchronously, so it is not counted as part of the "pipeline"). Note that some stage may take more than 1 cycles. Here are some of the abbreviations:

BTB = Branch Target Buffer
UOP = Micro Operations
ROB = Reorder Buffer
RF = Register File
BR = Branch
6R3W = 6 reads and 3 writes at the same time
3R2W = 3 reads and 2 writes at the same time
I$ = Instruction Cache (You can view it as instruction memory)
D$ = Data Cache (You can view it as data memory)
INT = Integer
iMul, iDiv, Int2FP = Integer Multiply, Integer Divide, Integer-to-Floating-Point
FP, FPDiv, FMA = Floating-Point, Floating-Point Divide, Floating-point Multiply and Add

Please answer the following questions.

(o) (5 pts) How many operations can be started in the Execute stage at the same time? What kind of operations? Why?

(p) (5 pts) Where does branch prediction occur in this pipeline? What happens if a branch is mispredicted? How many pipeline bubbles would it cause? Why?

(q) (5 pts) Suppose this is a dynamically scheduled pipeline. Please point out where the reservation stations are in the figure and how reservation stations work for this pipeline.

(r) (5 pts) Can data hazards occur in this pipeline? Please discuss which data hazards occur in this pipeline and how to reduce the pipeline bubbles caused by the data hazards?

(s) (5 pts) Please describe how interrupts/exceptions are handled by this dynamically scheduled pipeline where instructions can be executed out-of-order and multiple interrupts/exceptions caused by different instructions. [Hint: "Commit".]

Fetch
(4 cycles)
Decode
and
Rename
Rename
and
Dispatch
Issue
Register Read
Execute
Memory
Writeback
I$
Backing
Predictor
Fetch
Buffer
Decode Unit and
Rename Logic
Rename Logic
and Dispatch
ROB
Commit
Writebacks, Wakeups,
and more
Mem. Issue Queue
ALU Issue Queue
FP Issue Queue
Physical INT
RF (6R3W)
Physical FP
RF (3R2W)
x2
x2
x2
x3
MemCalc
ALU
BR/ALU
iMul
iDiv
Int2FP
FP2Int
FPDiv
FMA
D$ Shim
SAQ
SDQ
LAQ
D$
To Int/FP
RF
To Int
RF