

Assignment 1 Handout

Scalable AI: Bridging Theory, Understanding, and Practice

EE 290 / 194 · Spring 2026

Released	January 27, 2026
Due	February 10, 2026
Where you work	Course-provided H100 instance (8 × H100 per group)
Grading split	Part A (warm-up notebooks): 40% Part B (MFU deep dive + PR): 60%
What you submit	Completed notebook artifacts (Part A) + profiling + report + PR (Part B).
How To Submit	Please make sure to submit your assignment (notebooks, profiling results, report) on Gradescope.

Part A (40%): Finish the two notebooks

1. Notebook 1: Arithmetic Intensity (environment + roofline thinking).

Purpose: get familiar with the instance you will use for the course and with the compute-vs-bandwidth mental model. Follow the notebook and produce its required artifact(s) (e.g., `submission/answers.yaml`).

2. Notebook 2: Mixtral MFU Optimization Challenge.

Purpose: practice the workflow for improving MFU. You *can* brute-force configs, but it usually wastes time. The expectation is: **profile** → **form a hypothesis** → **make a focused change** → **re-measure**. Save evidence (Nsight traces, logs, configs) to support your claims and show how you optimized the Mixtral model.

Part B (60%): MFU on real training workloads

After Notebook 2, the real assignment begins: an open-ended systems investigation on real training/SFT workloads.

Your job:

- Choose a training/SFT stack and run workloads end-to-end. For example, you may start from NeMo AutoModel or NeMoRL with any backend (DTensor-based or Megatron-Core-based).
- Please select Nemotron-Nano-V3 as your target model.
- If your framework does not support Nemotron-Nano-V3, make sure you integrate support for it. If it does, you can use the existing support and improve upon it.
- For the framework you have chosen, measure various efficiency metrics (throughput and MFU, etc.), then profile and explain where performance is being lost.
- Diagnose bottlenecks with evidence, then design and implement interventions (anything reasonable, including writing CUDA kernels, reimplementing abstractions, etc.). You are welcome to go as grand as you want.
- If you can improve performance and MFU over the baseline behavior in the repo/framework, open a **PR** with reproduction steps.
- If the improvement is meaningful and reproducible, we will help publicize it.

What we care about

Your report and presentation must explain the path from observation to fix:

symptom → bottleneck (with evidence) → intervention → measured effect

We will ask you to defend why you believed something was the bottleneck, what evidence supported that belief, and why your change should help from first principles. Generally, had you ****only**** done a exhaustive random search, you probably would not be able to write a good report here.

Deliverables

- **Part A:** required notebook artifacts per the notebook checklists.
- **Part B:** baseline and improved commands/configs, plus scripts/notebooks used for measurement.
- **Profiling evidence:** traces/summaries that support your diagnosis (include enough context to reproduce).
- **Pull request:** before/after MFU numbers and clear reproduction steps.
- **Report (PDF) + short presentation.**

Required report structure

(Keep it concise: **1–4 pages** plus an appendix if needed.)

1. Setup + exact reproduction steps for baseline.
2. Baseline MFU/throughput and what “current OSS best” looks like.
3. Profiling results: key traces and what they imply.
4. Diagnosis: a few evidence-backed hypotheses.
5. Interventions: what you changed and why it should help.
6. Results: before/after + good ablations explaining why it helps.
7. Reproducibility notes: seeds, caveats, end-to-end run.

Hints for interventions

Examples (not exhaustive): parallelism strategy, communication overlap, batch/microbatch choices, precision/quantization, activation checkpointing, kernel fusions, optimizer settings, data/input pipeline.

Grading rubric (100%)

Component	Weight
Part A: Notebooks 1–2	40%
Part B: Measurement setup + baseline reproduction	10%
Part B: Diagnosis with profiling evidence	20%
Part B: Intervention engineering + PR quality/reproducibility	20%
Part B: Report + presentation	10%

Additional notes on Negative Results

You can earn full credit without “winning” the MFU wars, but you cannot earn full credit with ungrounded random search. Even if you do not improve performance, you can still earn full credit for a good diagnosis and interventions that you attempted, even if they did not help improve over baseline. We will evaluate you on the quality of your investigation, the systematic nature of your experiments, and the clarity of your thinking.