2026 Spring

Systems Reading Group

RG 26Spring Committee

2026.04.07

Agenda

- **Reading Group**
  - review
  - mission
- What's new?
  - talk organization
  - paper source
  - ...
- Advice for reading and presenting

Best presentations for last semester

TOPIC	PRESENTERS	TITLE
RL	Wei Gao (Alibaba ROLL)	ROLL: An Efficient and User-Friendly Scaling Library for Reinforcement Learning with Large Language Models
Training	Chenhan Wang, Luofan Chen	TrainVerify: Equivalence-Based Verification for Distributed LLM Training
Attention	Ping Gong, Xin Ren	Kimi Linear: An Expressive, Efficient Attention Architecture
Serving	Mingxuan Liu (西工大)	Jenga: Effective Memory Management for Serving LLM with Heterogeneity
RDMA	Yicheng Zhang	Spirit: Fair Allocation of Interdependent Resources in Remote Memory Systems
RAG	Chao Bi	HedraRAG: Co-Optimizing Generation and Retrieval for Heterogeneous RAG Workflows

Congratulations!

Best services for last semester

COORDINATION

Zhihui Chen

- Wechat notifications
- GitHub page updating
- Rating

MEDIA

Ouxiang Zhou

- Zhihu publishing
- Video clipping
- Bilibili uploading

ON-SITE SUPPORT

Ruibo Liu & Ouxiang Zhou

- Fruit
- Snack
- Offline & Online projection

Thank you for keeping RG running every week!

New committee menbers for this semester

COORDINATION

Chizheng Fang

- Wechat notifications
- GitHub page updating
- Rating

MEDIA

Yuzhe Li

- Zhihu publishing
- Video clipping
- Bilibili uploading

ON-SITE SUPPORT

Mulong Li & Shen Fu

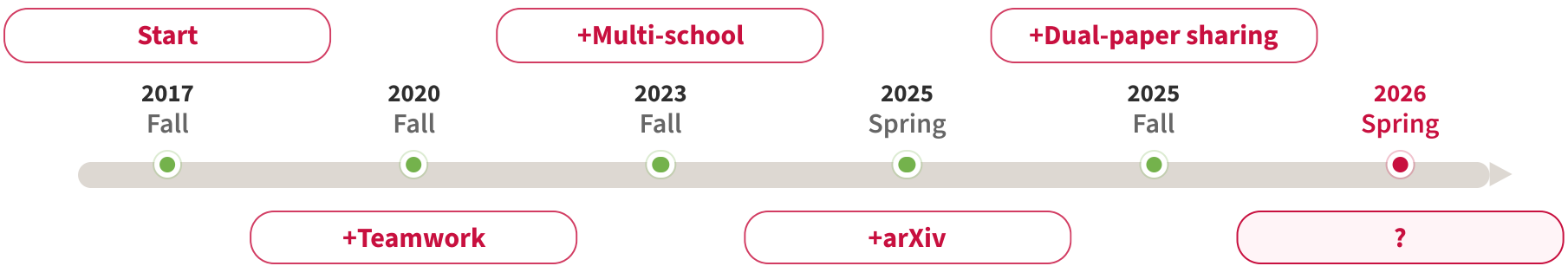
- Fruit
- Snack
- Offline & Online projection

Let's involve newbee members for the RG committee :)

Mission of reading group

- **Understand** and **keep up** with fast-moving systems research
- Learn how to do **solid, high-quality** systems research
- Polish the **soft skills** behind good research
  - research taste & critical thinking
  - presentation & writing
  - communication
  - ...

Roadmap of reading group



NEW PAPER POLICY

Topics focused, sources diverse.

to focus on one topic for several weeks at a time,
exploring it in both breadth and depth

Topic-based organization for this semester

Topic	时间	细化技术点与场景（包括但不限于）
模型基石：Attention 和 MoE	第 1-4 周	attention, MoE, long-context, KV cache reuse / migration, ...
执行基础：算子与 xPU	第 5-7 周	算子优化, 通算融合, Cuda/Tensor Core, 计算图, CuTeDSL, ...
基础系统：数据链路, 存储, 调度	第 8-10 周	RAG, agent memory, checkpoint, distributed storage, data pipeline, ANN, ...
新的变种：模型与硬件的新架构	第 11-13 周	linear attention, engram, 超节点, ...

论文出处 MLSys领域发展较快，限制OSDI和SOSP的论文无法满足日益增长的学习需求。
本学期不再限制具体出处。只需报名时经组织者确认，质量过关即可。

Arrangement

- When & where
  - **Time:** 19:00 - 21:00, 每周二
  - **Offline:** 信智楼 A707（视情况调整）
  - **Online:** #腾讯会议：840-4446-8248
- Entry points
  - **Schedule:** GitHub Page (https://adsl-rg.github.io/2026_spring.html)
  - **Call for talks:** Lark Doc (https://ncngj7vcrz76.feishu.cn/wiki/QG8fwxmhjiUl4PkEZf2cT0sDnGd)

时长要求 对于一篇论文，完整讨论: 45 - 50 分钟，分享: 30 - 35 分钟，建议30-40页slides。

Agenda

- Reading Group
- What's new?
- **Advice for reading and presenting**

Advice for reading a paper

- The three-pass approach (Srinivasan Keshav from University of Cambridge)
  - **1st pass:** get a bird's-eye view
  - **2nd pass:** grasp the core content
  - **3rd pass:** rethink the work and try to recreate the logic

Advice for reading a paper

- The three-pass approach (Srinivasan Keshav from University of Cambridge)
  - **1st pass:** get a bird's-eye view
  - **2nd pass:** grasp the core content
  - **3rd pass:** rethink the work and try to recreate the logic
- Reading with new tools
  - Chinese-friendly for arXiv papers: https://hjfy.top

Advice for reading a paper

- The three-pass approach (Srinivasan Keshav from University of Cambridge)
  - **1st pass:** get a bird's-eye view
  - **2nd pass:** grasp the core content
  - **3rd pass:** rethink the work and try to recreate the logic
- Reading with new tools
  - Chinese-friendly for arXiv papers: https://hjfy.top
  - vibe reading with prompts:
    - 假设我是一个很多前年计算机专业毕业的老奶奶，我想学习一下最近的这篇论文。请用傻子都能懂的语言详细给我讲一下这篇文章做了什么、怎么做的，禁止篡改文章原意，不要离文章核心。
    - 帮我用通俗易懂的语言解释一下这篇论文关注的场景、要解决的问题、现有**SOTA baseline**为什么无法解决、新的发现与分析、遇到的挑战、提出的技术方案、用到的实验平台与场景以及最后的实验效果。

Advice for presenting

- Primary focus: really understand the paper
  - What problem and scenario does it concern?
  - What are the SOTA baselines, and what is missing there?
  - What are the core insights and techniques?
  - What are the key challenges?
  - What do we actually learn from the evaluation?

Advice for presenting

- Primary focus: really understand the paper
  - What problem and scenario does it concern?
  - What are the SOTA baselines, and what is missing there?
  - What are the core insights and techniques?
  - What are the key challenges?
  - What do we actually learn from the evaluation?
- Tips
  - 1 or 2 minutes per slide is good
  - Too much text is a warning sign
  - Please do rehearsals offline
  - Clearly give some concluding sentence per slide
  - Before presenting evaluation results, give the motivation per experiment

Advice for presenting

RESOURCE 1

Preparing a talk / Giving the talk

Good for structure, timing, and delivery basics.

Advice for presenting

RESOURCE 1

Preparing a talk / Giving the talk

Good for structure, timing, and delivery basics.

RESOURCE 2

Oral presentation advice / How to give a bad talk

Good for learning from common mistakes.

Advice for presenting

RESOURCE 1

Preparing a talk / Giving the talk

Good for structure, timing, and delivery basics.

RESOURCE 2

Oral presentation advice / How to give a bad talk

Good for learning from common mistakes.

TIPS Feel free to use fancy LLM tools.

RG 2026 Spring

Please sign up!

Discussion and QA

One More Thing

Considering the fast iteration of MLSys and LLM tools,
we **encourage fancy LLM tool sharing**
within 5-10 minutes after discussion.

RG 2026 Spring

Please sign up!

Discussion and QA