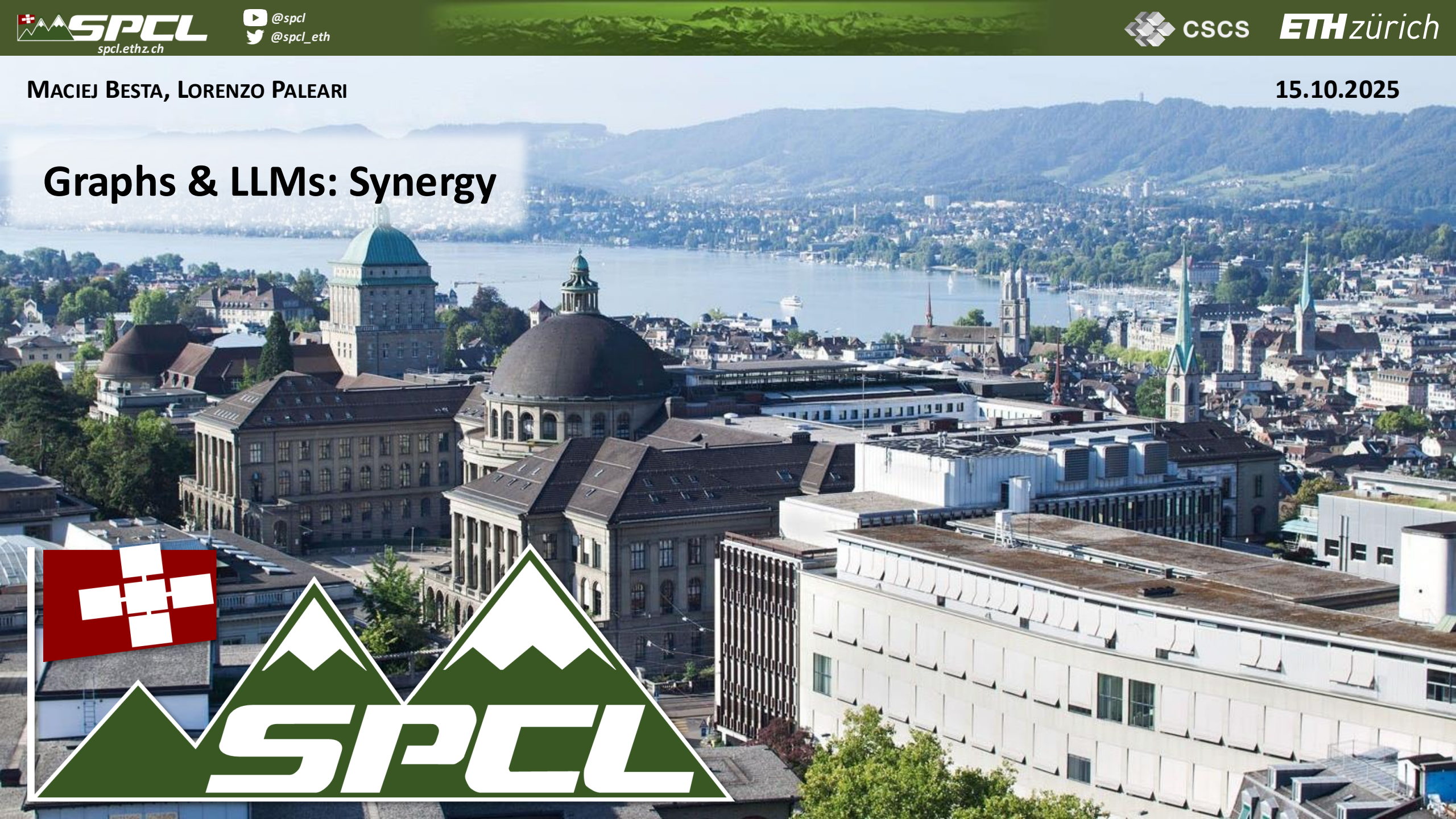
Maciej Besta, Lorenzo Paleari

15.10.2025

Graphs & LLMs: Synergy



KV Caching -> LoRA

- **Observation**: During inference, the model builds a KV cache (K,V) capturing contextual activations for each token --> (Ex. ChatGPT)
- **Problem**: These caches are discarded after use (Too big to be stored for each chat and user)
  - Waste of rich latent information
  - Limits contextual continuity --> New chat without this information's (need to recompute KV Cache and store it (?))
- **Idea**:
  - Implicit context compression --> the model "remembers" without storing information's/KV-Cache

Problems

- How to make the model learn at this point?
  - Training – We need data that we are missing, and it is not available
  - LoRA – Can be used, but needs a little data to work
  - The problem can be directly solved → mimic GPTQ paper

- GPTQ paper
  - They solve the problem of efficient quantization by solving the underlying equation problem.
  - This allow very fast and efficient computation
  - We can follow same step and expand the math to our problem

$$\text{argmin}_{\widehat{\mathbf{W}}} \|\mathbf{W}\mathbf{X} - \widehat{\mathbf{W}}\mathbf{X}\|_2^2.$$

$$w_q = \text{argmin}_{w_q} \frac{(\text{quant}(w_q) - w_q)^2}{[\mathbf{H}_F^{-1}]_{qq}}, \quad \boldsymbol{\delta}_F = -\frac{w_q - \text{quant}(w_q)}{[\mathbf{H}_F^{-1}]_{qq}} \cdot (\mathbf{H}_F^{-1})_{:,q}.$$

$$\mathbf{H}_{-q}^{-1} = \left(\mathbf{H}^{-1} - \frac{1}{[\mathbf{H}^{-1}]_{qq}} \mathbf{H}_{:,q}^{-1} \mathbf{H}_{q,:}^{-1}\right)_{-p}.$$

Where to apply this

- First to System Prompt
  - Async
  - Can be slower
  - Bigger matrices possible
- Expand to Context
  - Sync
  - Faster
  - Little matrices