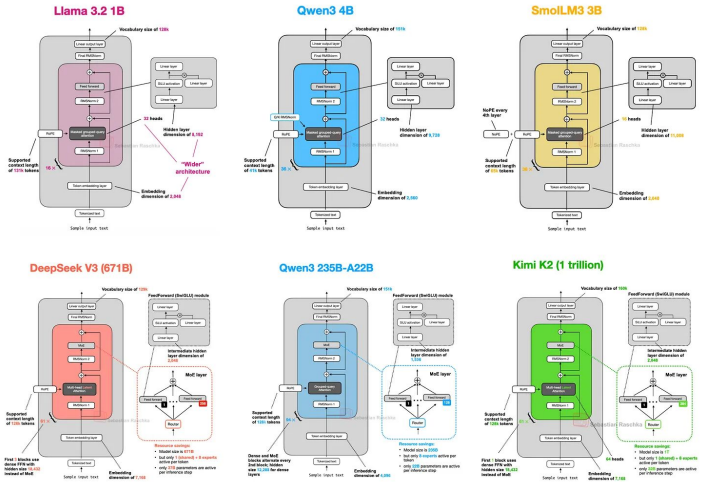
Distributed Inference

Boris Gans, Yousef Amirghofran, Lea Aboujaoude, Anže Zgonc, Matthew Porteous, Alp Arslan Baghirov

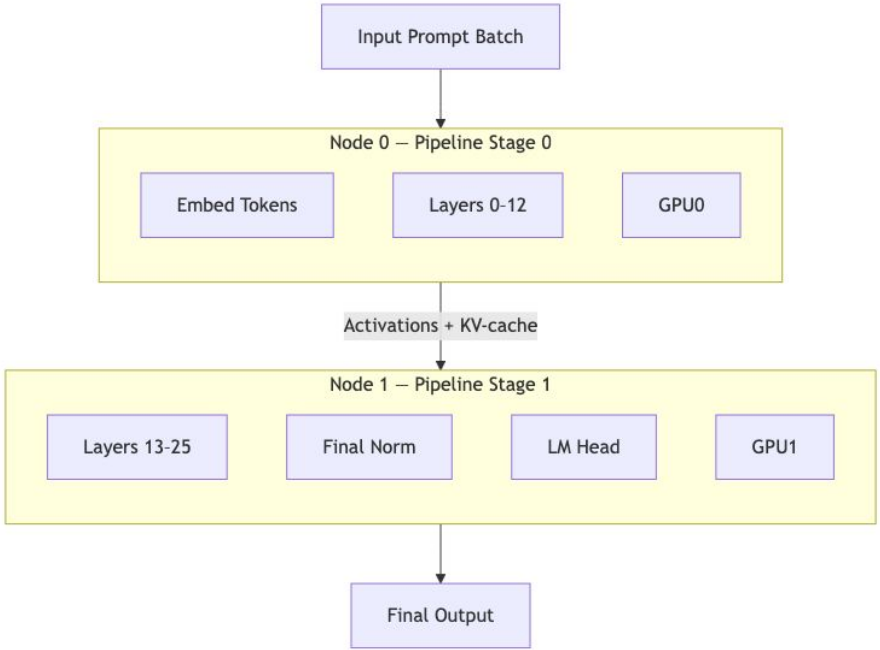
Problem & Impact

- Demand for large language models has increased exponentially
- More powerful models are generally bigger (number of parameters)
- They can’t be fit into one node
- We need to distribute the inference to these models
- This lets us take advantage of more powerful models



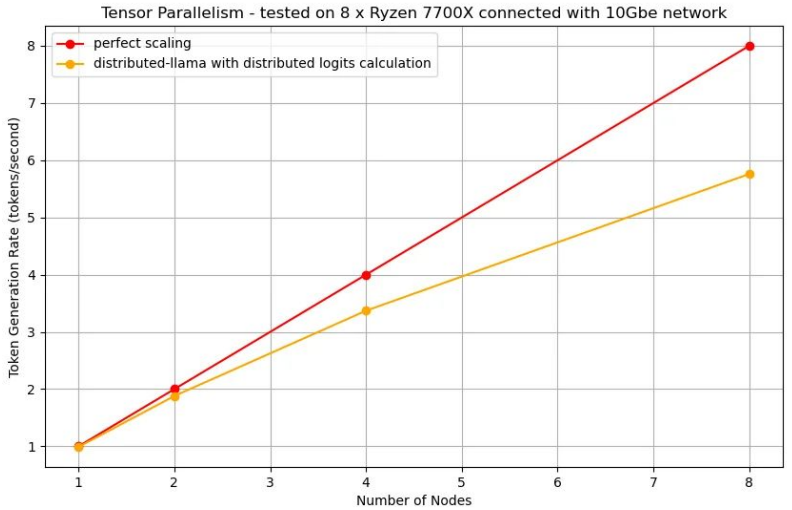
Approach & Prototype

- Two-stage pipeline-parallel deployment of OpenLLaMA 3B v2
- Containerized execution: each rank runs inside an Apptainer environment
- Performed three experiments:
  - Strong Scaling: Fixed workload (5 prompts), varied number of nodes
  - Weak Scaling: Increase workload and nodes proportionally
  - Batch-size Sweep: Vary batch size (1, 2, 4) to observe model behavior



Scaling & Profiling

- Present system does not scale nicely
  - Limited VRAM forces heavy disk offload
  - No NVLink → forces traffic over PCIe + network
- Nsight unavailable on the cluster → profiling relies on in-application timers



EuroHPC Targets & Resource Requests

Project Targets

- Build scalable multi-node LLM inference pipeline
- Enable deep pipeline parallelism & activation offloading
- Integrate GPU profiling (Nsight)
- Scale to 7B–13B models on EuroHPC GPUs

Why EuroHPC

- Multi-GPU nodes with NVLink
- InfiniBand for inter-node activation transfer
- A100 40GB needed for large-model partitions
- Access to Nsight profiling tools

Resource Request

- ~4,000 GPU node-hours total
- Includes debugging, containerization, pipeline scaling, profiling
- Requires 2–16 GPU jobs for scaling studies

Risks, Milestones, and Needed Support

Risks

- **Model too large for GPU memory** → solved with checkpointing & quantization
- **Pipeline imbalance** → solved with automated layer-to-GPU mapping
- **NCCL / topology issues** → mitigated with profiling + topology-aware placement
- **Profiling tool availability** → fallback to internal timers (unlikely on EuroHPC)

Needed Support

- Access to **Leonardo Booster** (A100 NVLink + InfiniBand)
- Availability of **Nsight Systems / Nsight Compute**
- Reliable multi-node scheduling for 4–16 GPU jobs
- Assistance with cluster **NCCL configuration** if needed

Key Milestones: