SC24 Atlanta, GA | hpc creates.

Tutorial. Nov 18, 2024 1:30-5 pm

Distributed Deep Learning on GPU-based Clusters

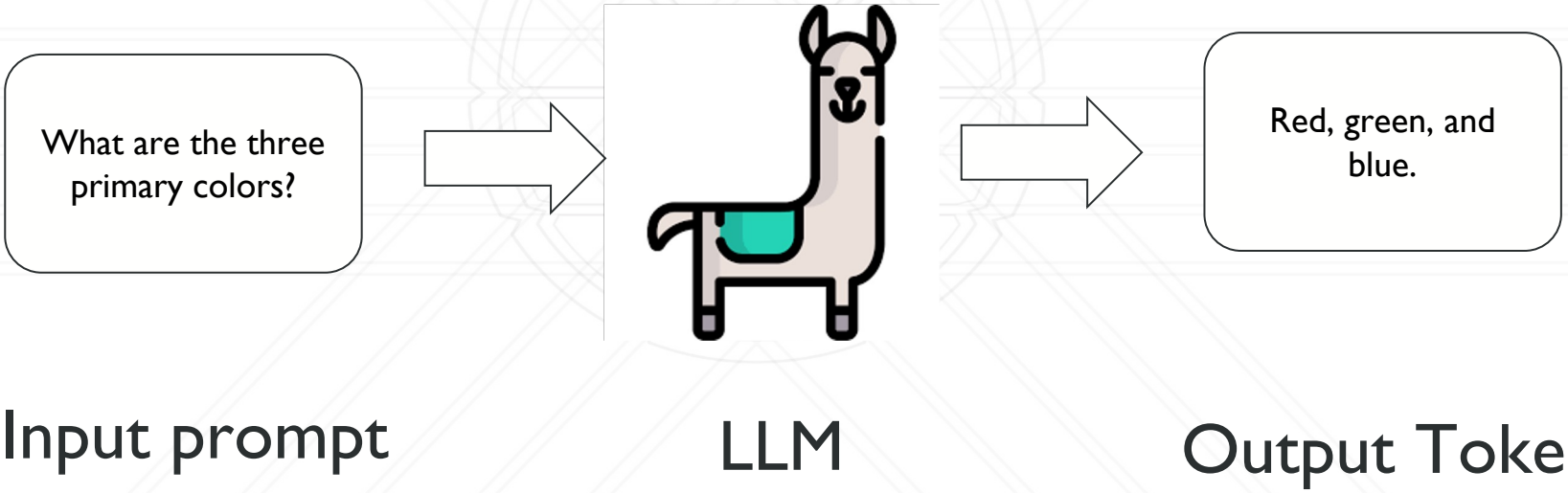
Abhinav Bhatele, Siddharth Singh, Prajwal Singhania
Department of Computer Science

PARALLEL SOFTWARE AND SYSTEMS GROUP

UNIVERSITY OF MARYLAND

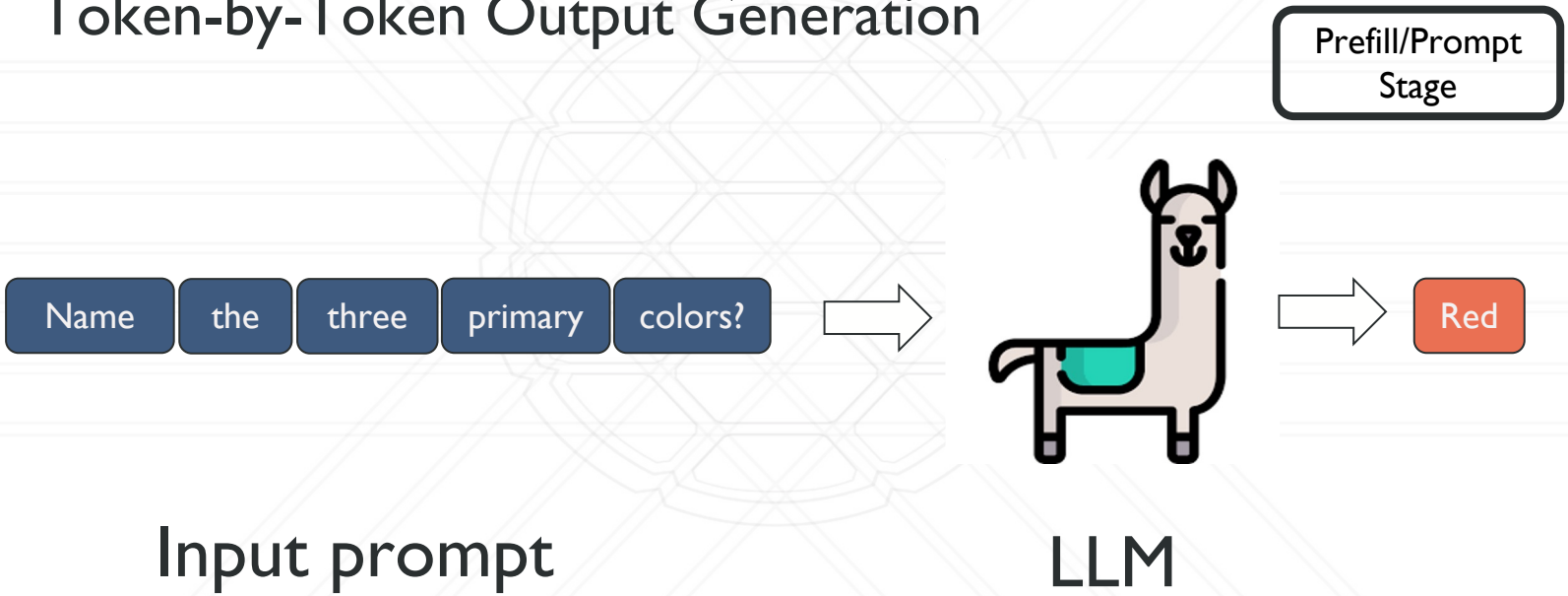
Inference

- Generating outputs from a trained language model



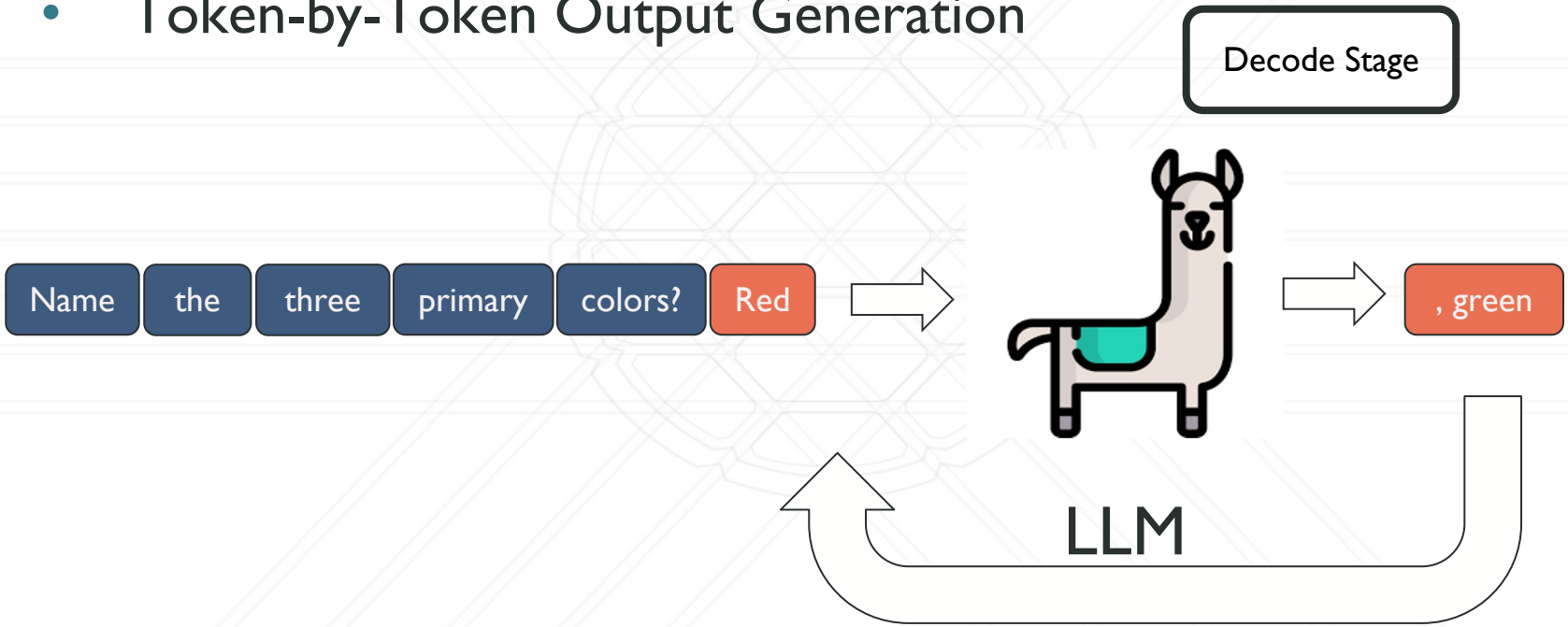
Autoregressive LLM Inference

- Token-by-Token Output Generation



Autoregressive LLM Inference

- Token-by-Token Output Generation



Types of Inference

- Online Inference
  - Generating outputs in real time as user input is received.
  - Ex -, Claude
- Offline inference
  - Generating outputs on pre-collected data
  - Ex - Synthetic data generation, benchmarking

Inference Frameworks

- Why is inference different from training to require separate optimizations/frameworks?
  - **Auto-regressive** nature of inference
    - Memory Bound!
  - No backward pass in inference
  - Different workload characteristics, eg. smaller batch size

Inference Frameworks

- What does an inference framework provide you?
  - Model Implementations with/without optimized kernels
  - Memory and request management

vLLM

LitGPT

Hugging Face

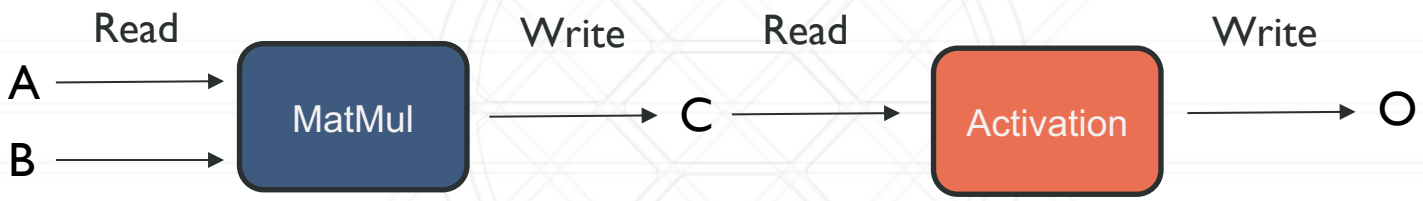
Hands-On: LitGpt Inference

- Part 1:

```
with axonn.models.transformers.parallelize():
    model = <declare hf transformers model>
```

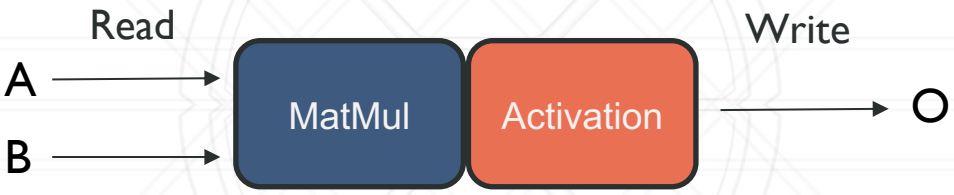
Torch Compile

- Kernel Fusion



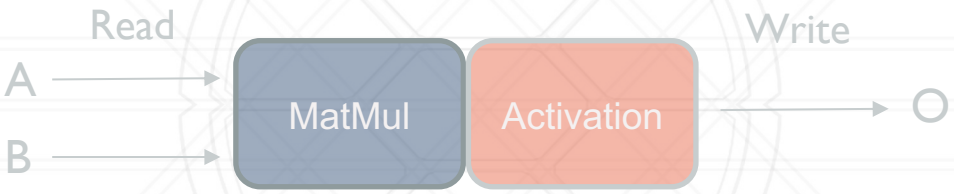
Torch Compile

- Kernel Fusion



Torch Compile

- Kernel Fusion



- CUDA Graphs: Create a computation graph to launch multiple kernels at once

Hands-On: Using Torch Compile

```
with axonn.models.transformers.parallelize():
    model = <declare hf transformers model>
```

Hands-On: LitGpt + AxoNN

```
with axonn.models.transformers.parallelize():
    model = <declare hf transformers model>
```

Running the code

```
sbatch --reservation=sc2024 infer_hf_axonn.sh prompts.txt
```

Try adding your own prompts to this file

Running the code

```
sbatch --reservation=sc2024 infer_vllm.sh prompts.txt
```

Try adding your own prompts to this file