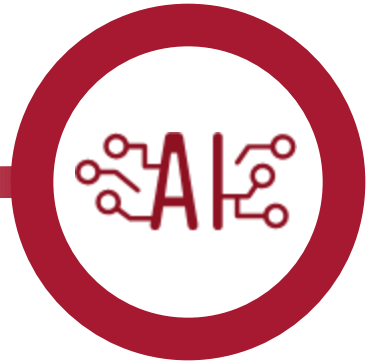
Per3S

AI for Science and Exascale

Large Language Models

The Rise of Data



Louis Douriez
ldouriez@ddn.com

Natural Language Processing is driven by Transformers

ChatGPT

Model: GPT-4

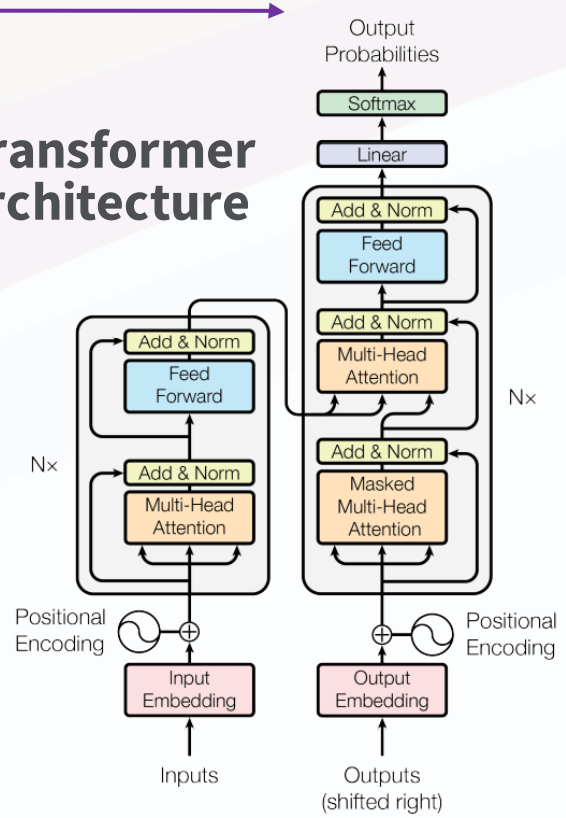
L Tell me a joke

Sure, here's a classic for you:

Why don't scientists trust atoms?

Because they make up everything!

Transformer architecture



Output

FFN

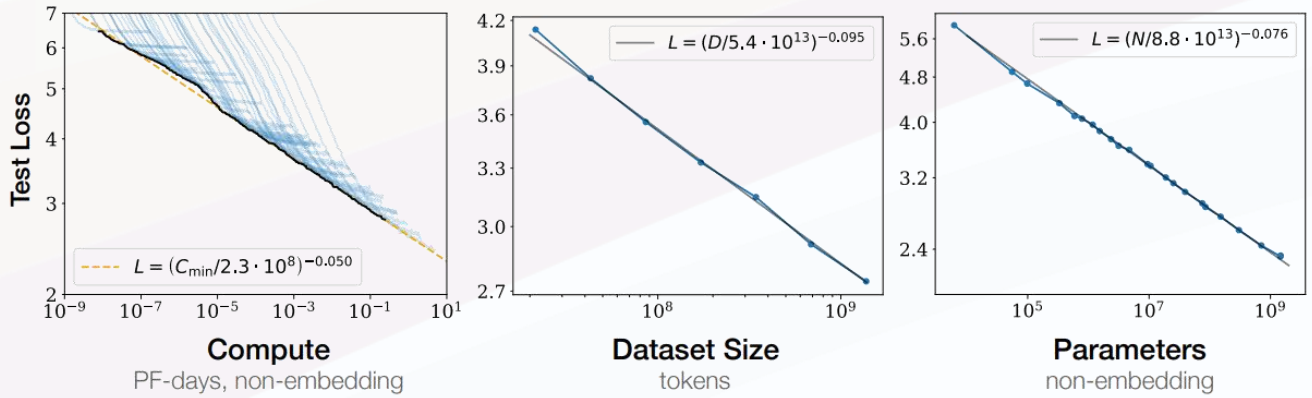
Add context

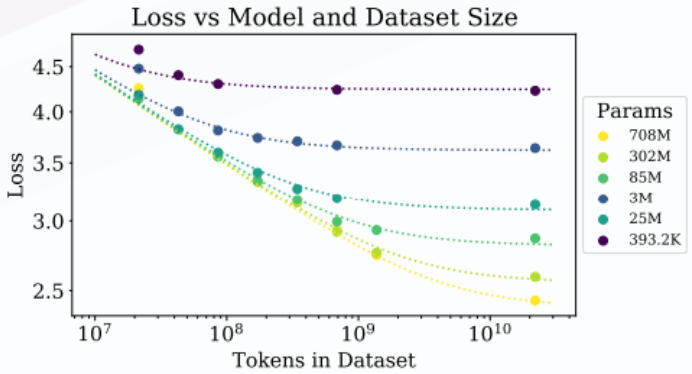
input

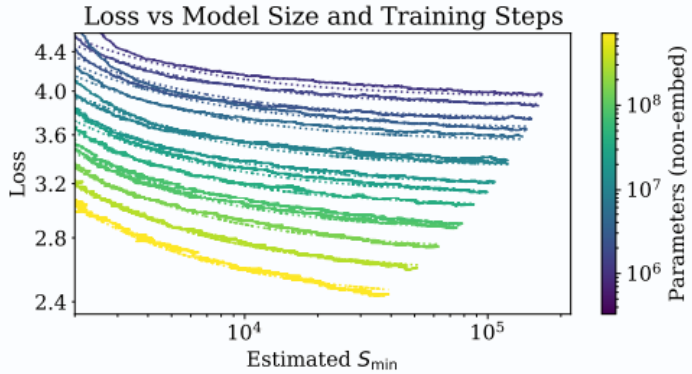
[ASH17] Ashish Vaswani, Noam Shazeer, Niki Parmar et al. *Attention Is All You Need*

Transformers drive Large Language Models (LLM)

More parameters equal better loss

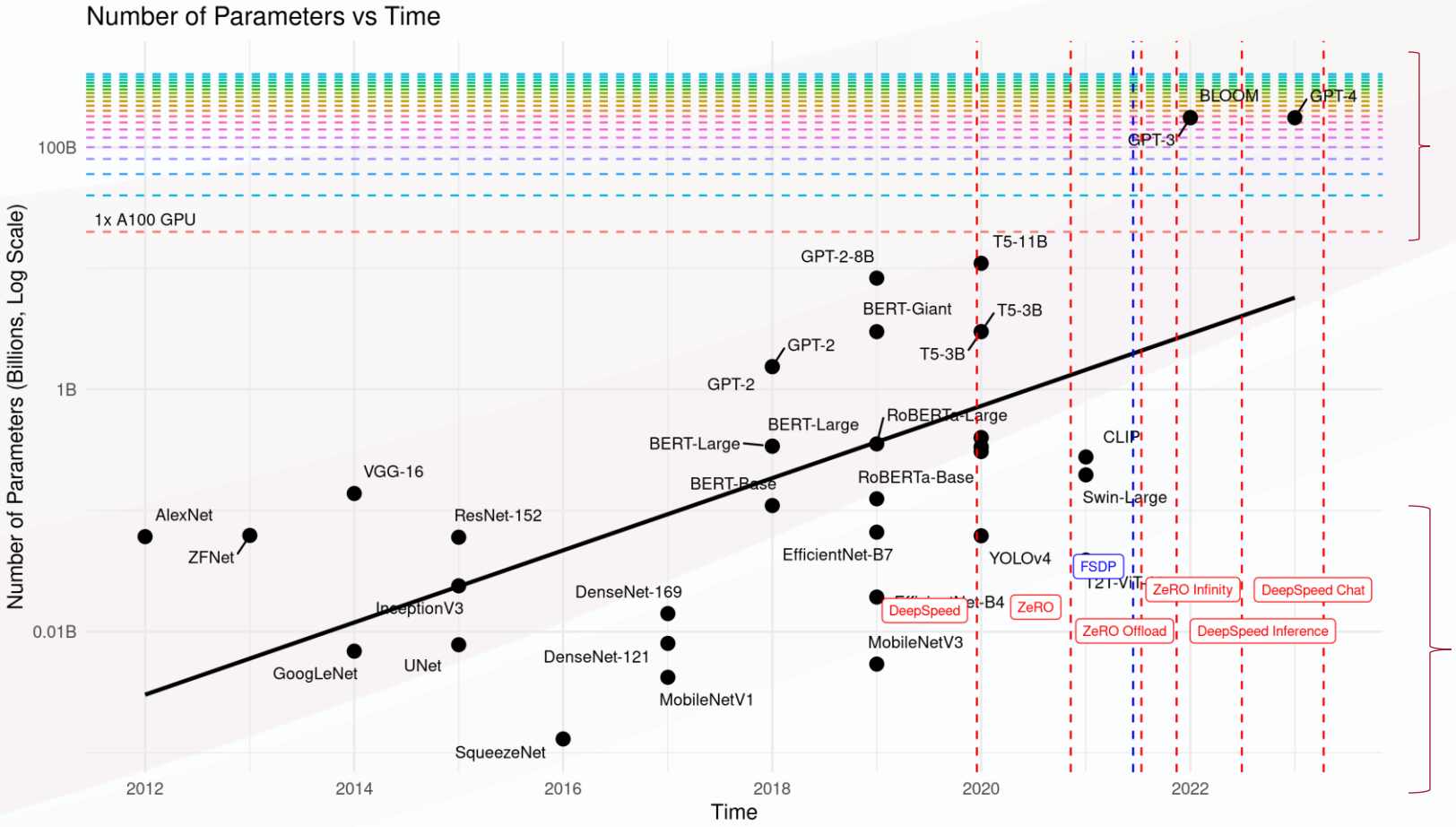






[KAP20] KAPLAN, Jared, MCCANDLISH, Sam, HENIGHAN, Tom, *et al.* Scaling laws for neural language models.

Large Language Models and the GPU memory wall



Model size did x1000 in 3 years
GPU memory did x5

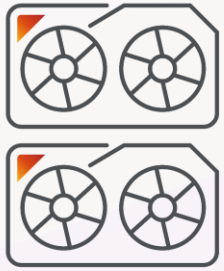
Software stacks (e.g: deepspeed) have been developed to handle the issue

- Better memory management (e.g: ZeRO)
- **Offloading (CPU/NVMe/Filesystem)**

LLM Offloading

What is model offloading?

Without model offloading, the model must fit in the aggregated GPU memory



With model offloading, the model is swapped in-&-out from the GPU to another form of storage (e.g local NVME)

NVIDIA DGX A100 is the world's first AI system built on the NVIDIA A100 Tensor Core GPU

The AI400X2 appliance is a fully integrated and optimized shared data platform

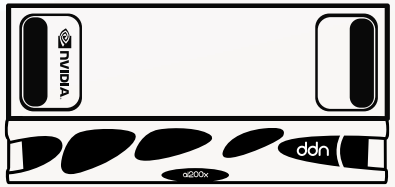
What if we offload a LLM on an AI400X2?

LLM Offloading Experiment

Using DeepSpeed ZeRO-Infinity

Setup

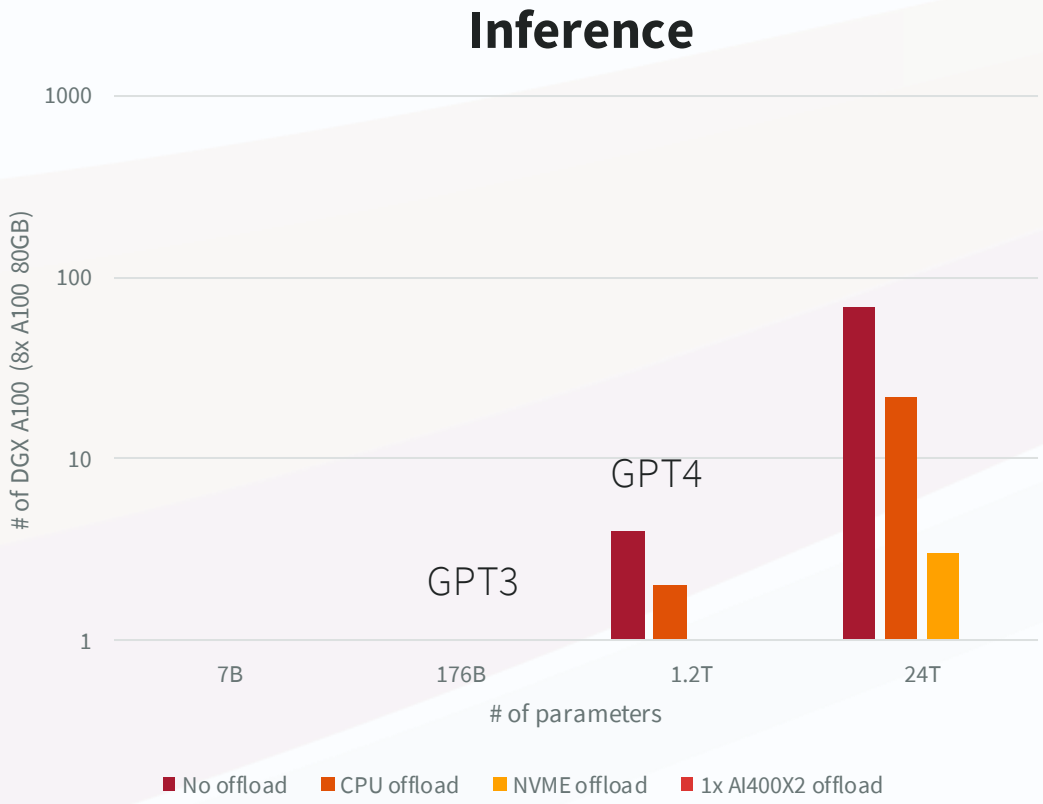
- 1x DGX A100
- 1x AI400X2
- 8x HDR200

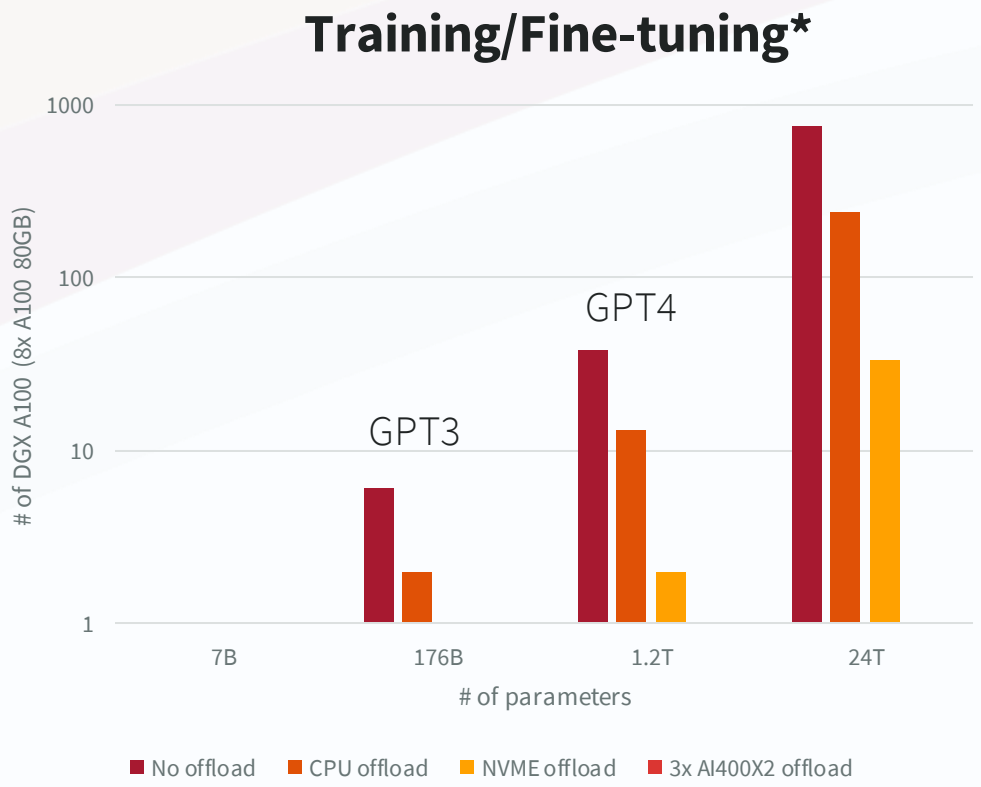


Offloading targets
GPU only
CPU offload (RAM)
NVME offload (local RAID)
AI400X2 offload

	Pre-trained (BLOOM)		Synthetic	
Model name	**BLOOM 7B1 par.**	**BLOOM 176B par.**	**BLOOM-mod-1 1.2T par.**	**BLOOM-mod-2 24.1T par.**
# hidden layers	30	70	960	4800
hidden-dim	4096	14336	10240	20480
# attention heads	32	112	16	16
Batch-size used	32	16	8	2

LLM Offloading Experiment - How many DGX needed (fp16)?



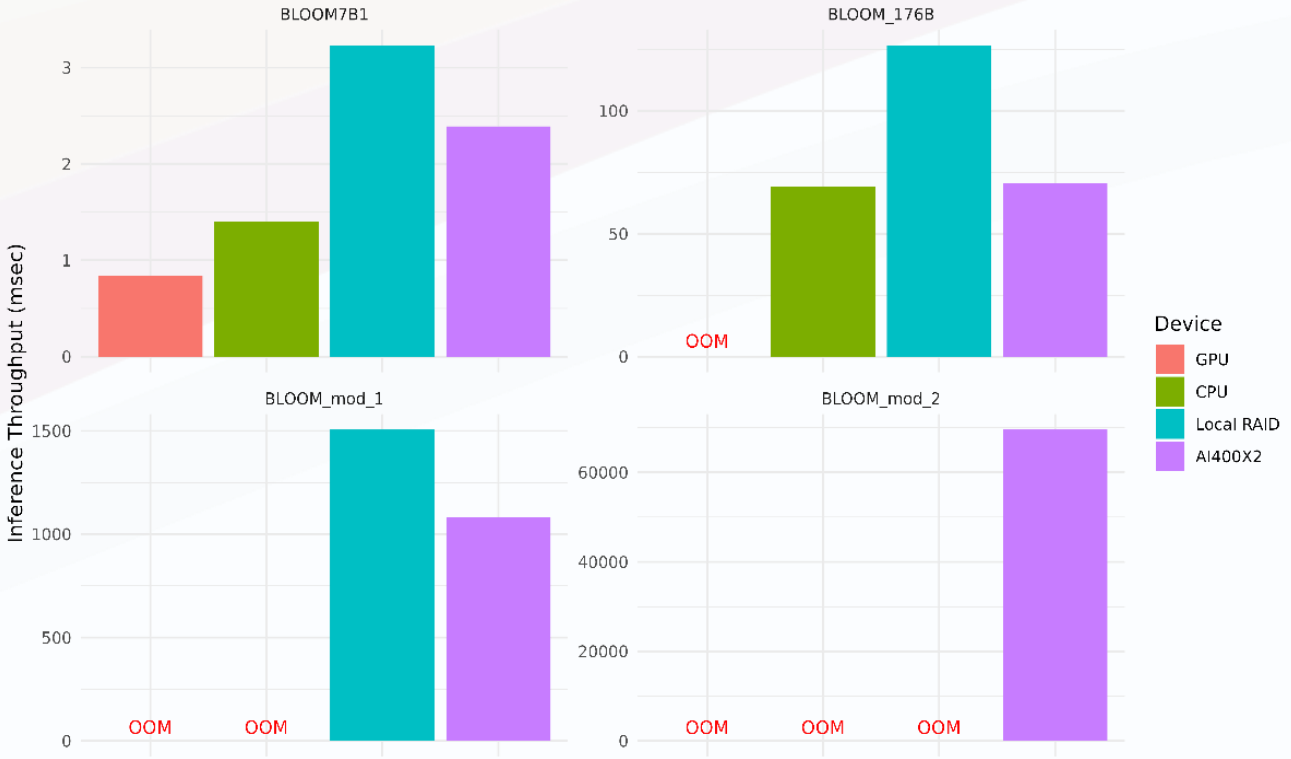


*extrapolation using $240 \times nl \times hd2$ [RAJ21] RAJBHANDARI, Samyam, RUWASE, Olatunji, RASLEY, Jeff, *et al.* Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning.

LLM Offloading Experiment – Impact on performance

Result for inference

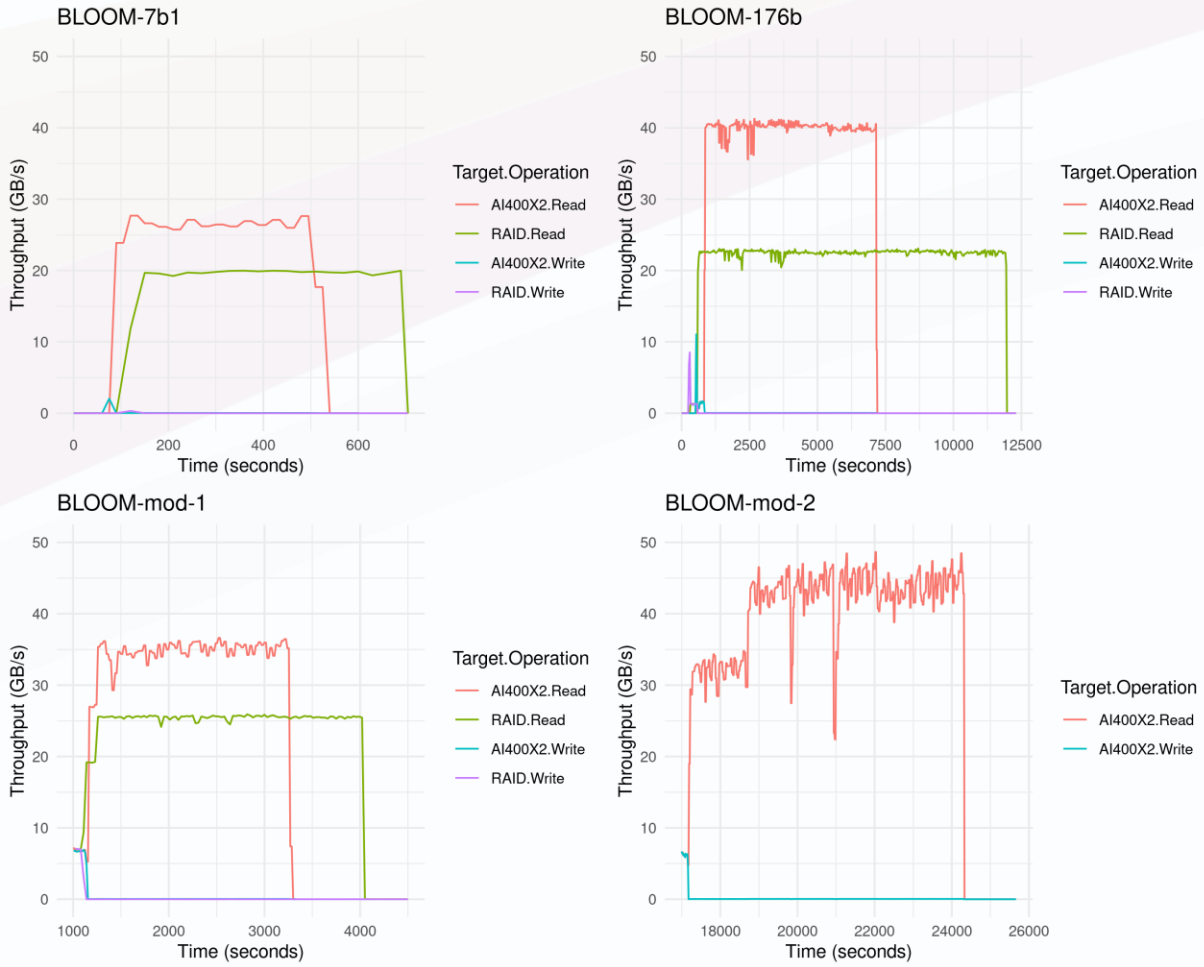
- The offloading on the AI400X2 outperform the local RAID of the DGX A100 for all test case (~2 times the throughput)
- The offloading on the AI400X2 equalize CPU offloading performance for GPT3 like models (<1%)
- The offloading on the AI400X2 can run inference on 24 Trillion parameters (x24 times params GPT4)



LLM Offloading Experiment – IO throughput

Result for inference

- The total amount of data transfered is the same between the local RAID and the AI400X2
- The IO throughput determines the performance
- Transfer is overlapped with computation. It is a throughput problem

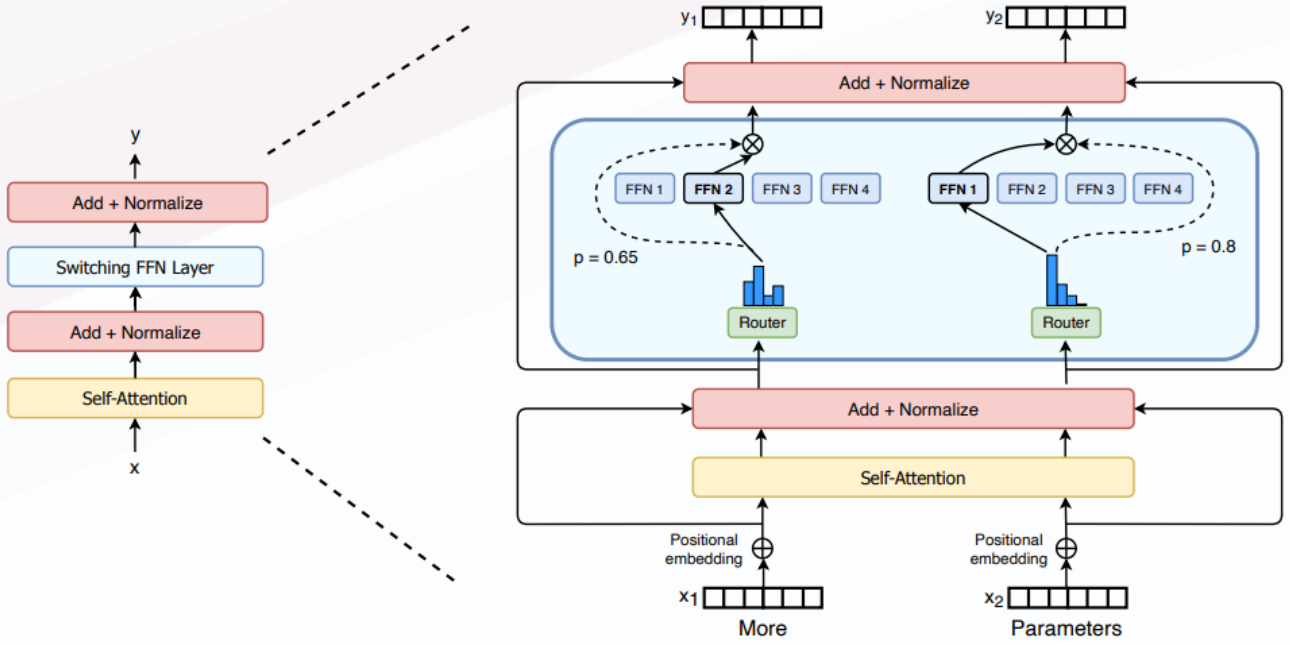


LLM Offloading – Use cases are not limited to inference

Large Language Models with low computation volume OR extremely large model

- Sparse models (MoE) (read/write)
- Inference (read only)
- Fine-tuning (read/write)
- Training on extremely large models that wouldn't fit on any system (read/write)

Sparse models decouple the size of the model with the amount of computation needed



Sparse model example: switch transformer

Thank You!

Questions?