Conclusion

- ZKPerf: a ZK-SNARK proving benchmark with crypto and ML tasks
- Most costly parts of ML circuits are nonlinearities (e.g., ReLU)
  - Need efficient lookups to prove larger models
- Circuit structure matters
  - One way: optimize row and column dimensions

uiuc-kang-lab/zkperf

@liliatangxy

liliat2@illinois.edu

DeServe: Building Decentralized Services for Collaborative Large Language Model Inference

Xiaoyuan Liu, UC Berkeley, 2024

Decentralized model serving

Problem scope

- Inference (~~fine-tuning~~, ~~training~~)
- Decentralized computation resource: federated / personal device
- Offline serving (latency insensitive)

Key questions

- Can decentralized serving gain cost advantages?
- How to optimize throughput in high-latency heterogeneous env?
- How to prevent fraud and protect computation integrity?

Decentralized model serving

Problem scope

- Inference (~~fine-tuning~~, ~~training~~)
- Decentralized computation resource: federated / personal device
- Offline serving (latency insensitive)

Key questions

- Can decentralized serving gain cost advantages?
- How to optimize throughput in high-latency heterogeneous env?
- How to prevent fraud and protect computation integrity?

Can decentralized serving **gain cost advantages?**

1. ChatGPT
   a. gpt-4o $5~$15 / 1M token
   b. **gpt-4o mini (batched - offline)** $0.075~$0.3 / 1M token

2. Mining profit (Aug 2, 2024)
   a. 1xRTX 4090 revenue: $0.82 per day (NEXA)

3. LLM serving requirement
   a. Llama-3-70b → mem: ~140G
   b. RTX 4090: 24G → at least 6 (ideally >=8)

Can decentralized serving gain cost advantages?

⇒ If you have 8x4090, mining for one day, you get $6.56
⇒ Serving 21.9 M token ⇒ 253 token/s

Take away:

if you build a serving framework with throughput higher than this bar, you can convince miners to run decentralized serving to earn more

Can decentralized serving gain cost advantages?

⇒ If you have 8x4090, mining for one day, you get $6.56
⇒ Serving 21.9 M token ⇒ 253 token/s

Take away:

if you build a serving framework with throughput higher than this bar, you can convince miners to run decentralized serving to earn more

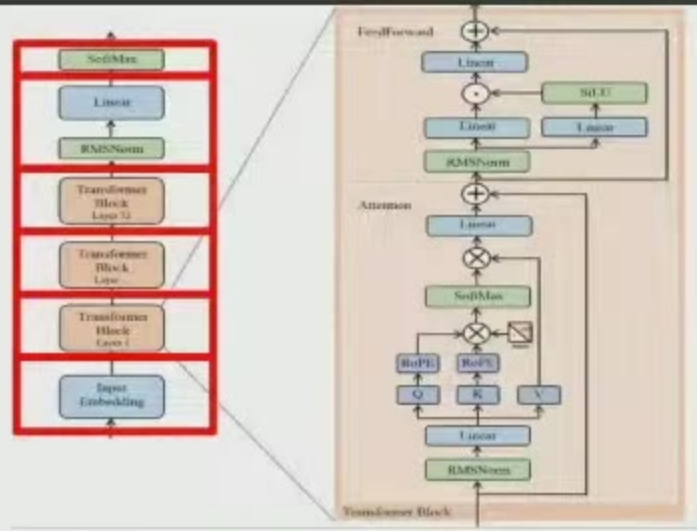
Can decentralized serving gain cost advantages?

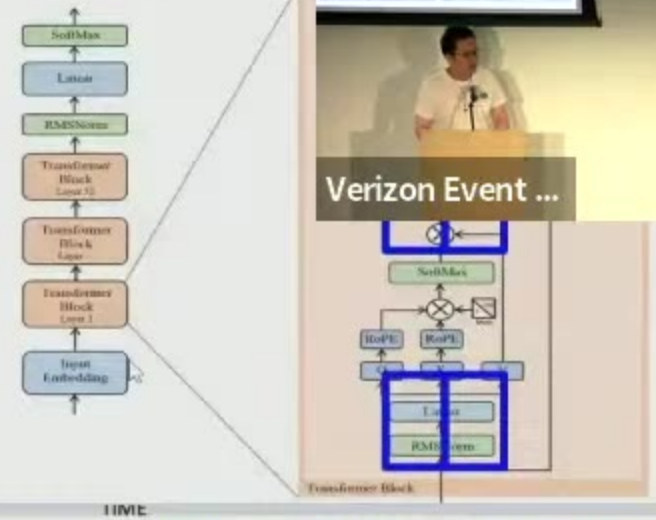
⇒ If you have 8x4090, mining for one day, you get $6.56
⇒ Serving 21.9 M token ⇒ 253 token/s

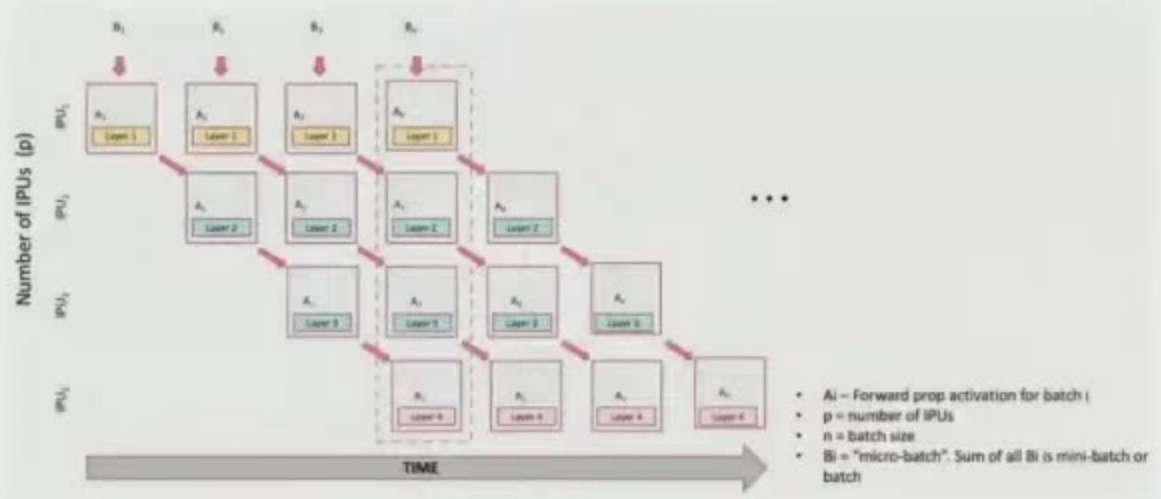
Take away:

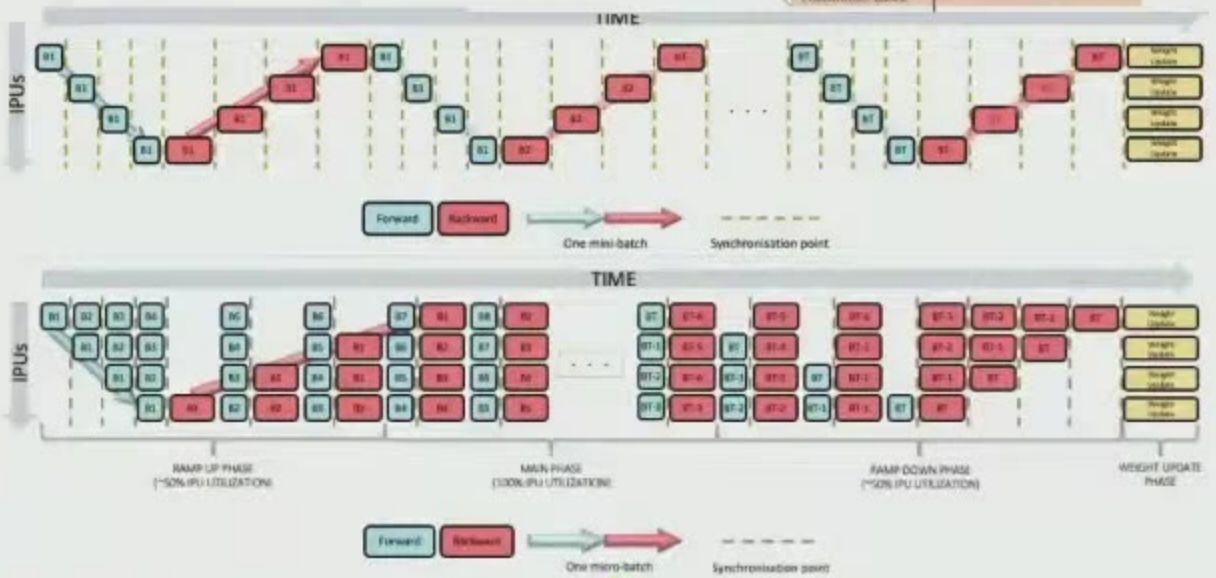
if you build a serving framework with throughput higher than this bar, you can convince miners to run decentralized serving to earn more

How to optimize throughput in **high-latency heterogeneous** env?









1. <u>tensor parallelism</u> 🚫 vs <u>pipeline parallelism</u> 👍
   a. Tensor parallelism requires frequent, high-bandwidth connection to be efficient (e.g. same machine)

Figure borrowed from: https://zhuanlan.zhihu.com/p/649756898 https://afmck.in/posts/2023-02-26-parallelism/

How to optimize throughput in **high-latency heterogeneous** env?

Applying optimization techniques:

+ Memory-aware (continuous) batching
+ Pipelining (in circle)
+ Flash attention
+ Paged attention
+ Optimized tensor serialization, etc…

⇒ **With 8x4090, 1071 token/s**
(bar: 253 token/s)

When latency increases:

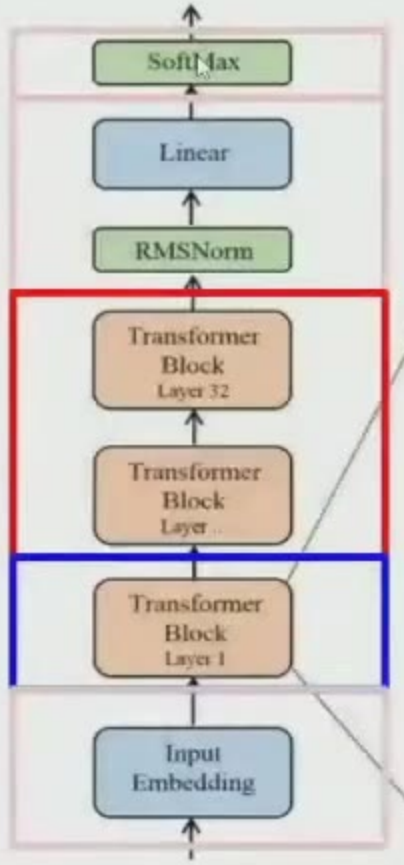
How to optimize throughput in **high-latency heterogeneous** env?

Heterogeneity with pipelining:

- Llama-3-8b: 32 layers
- Llama-3-70b: 80 layers
- Llama-3-405b: 126 layers

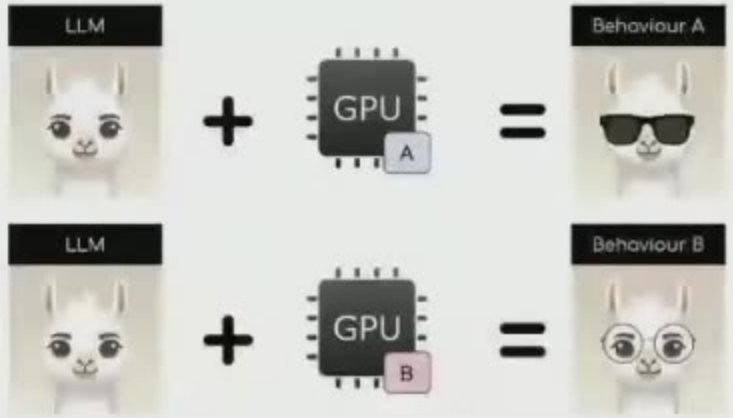
Faster device with more GPU mem

Slower device with less GPU mem



How to prevent **fraud** and protect computation **integrit**

Definition of "correctness": close to the "**standard**" (precise) result



Changing the GPU is changing the behaviour of your LLM.

Anis Zakari · Follow
12 min read · May 28, 2024

Different output (token prob), both **benign**

On the T4 side: "etc... This also means **you can trust** the output more since everything inside will be consistent across different runs!..."

On the A10G side: "etc... This also means **you can be more** confident when asking questions specifically related to topics covered within those texts..."

Reason: non associative float-point op

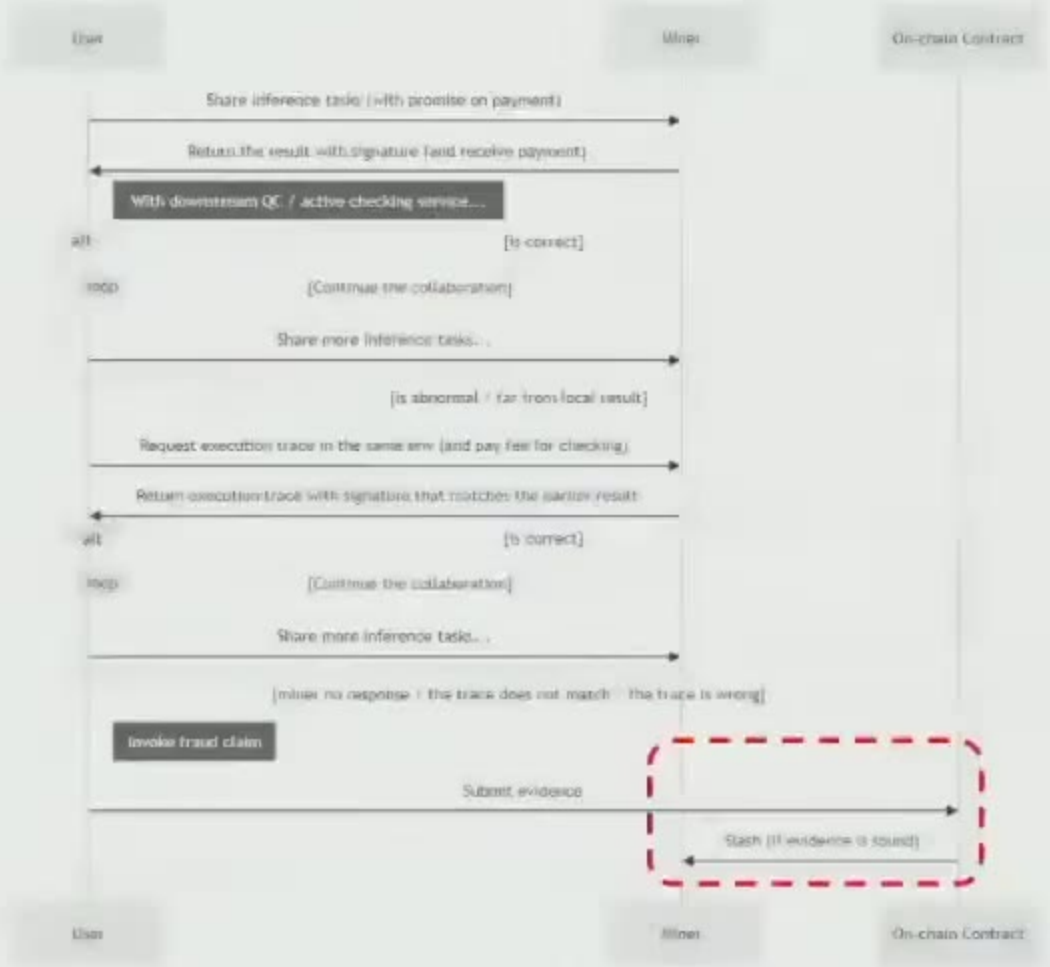
```
>>> sum([1e10] + [1e-10] * int(1e5))
10000000000.0
>>> sum([1e-10] * int(1e5) + [1e10])
10000000000.00001
```

> Can we make it the same / deterministic?

https://medium.com/@anis.zakari/changing-the-gpu-is-changing-the-behaviour-of-your-llm-0e6dd8dfaaae

How to prevent **fraud** and protect computation **integrit**

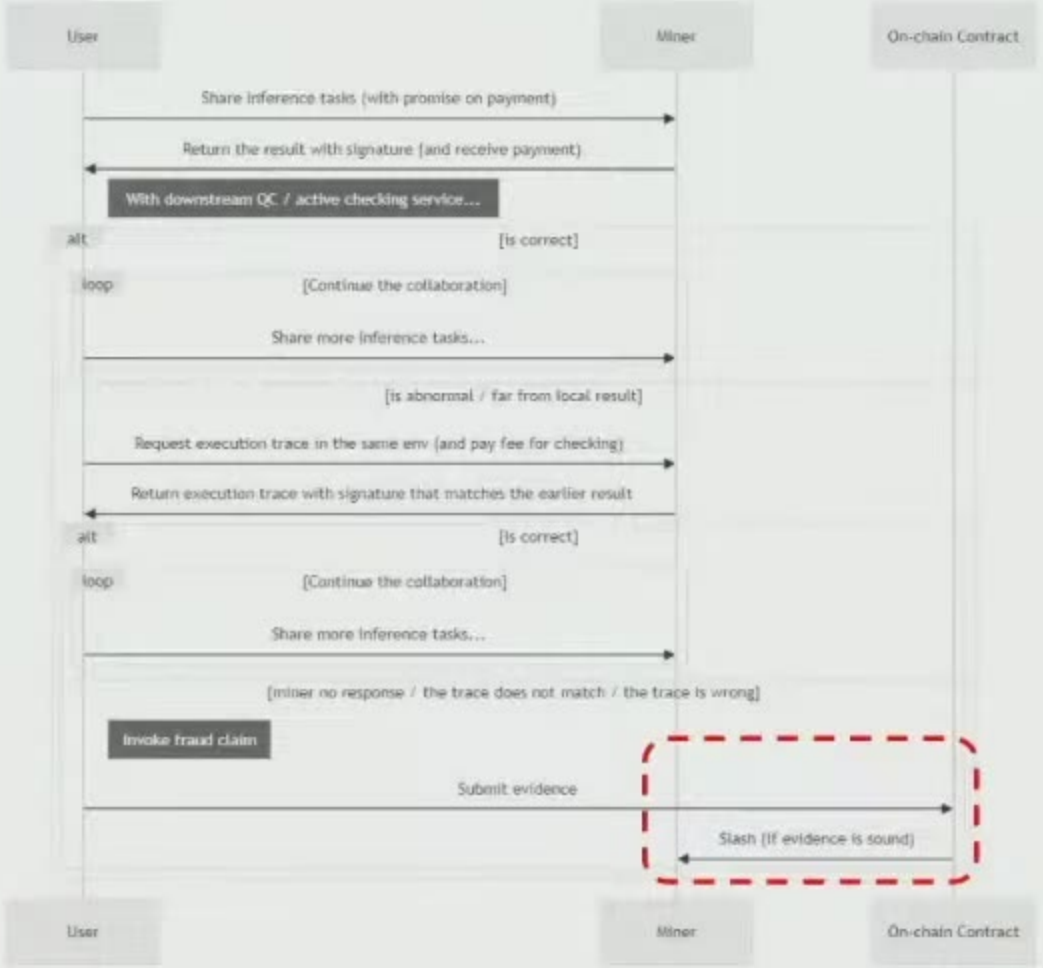
A simple optimistic **design**: trace checking with a threshold on tensor difference



Save on-chain computation:
Same design as step-proof

How to prevent **fraud** and protect computation **integrit**

A simple optimistic **design**: trace checking with a threshold on tensor difference



Save on-chain computation:
Same design as step-proof

Summit on Responsible Decentralized Intelli[illegible]

—— Future of Decentralization and AI

Thank you!

DeServe: Building Decentralized Services for Collaborative Large Language Model Inference

- Motivation: pricing model - [illegible]
- Serving efficiency optimization - model [illegible] through network
- Correctness & integrity - optimistic & deterministic serving

Xiaoyuan Liu, UC Berkeley, 2024

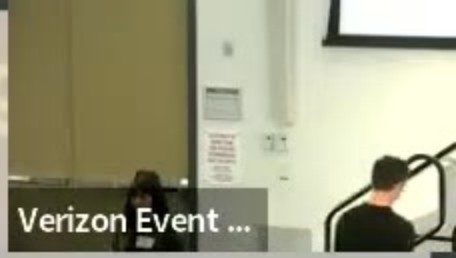
Powered by Zoom

Thank you!

DeServe: Building Decentralized Services for Collaborative Large Language Model Inference

- Motivation: **pricing** model - profit bar: 253 token/s/8x4090
- Serving **efficiency** optimization - model pipelining through network
- **Correctness** & **integrity** - optimistic & deterministic serving

Xiaoyuan Liu, UC Berkeley, 2024



Thank you!

DeServe: Building Decentralized Services for Collaborative Large Language Model Inference

- Motivation: **pricing** model - profit bar: 253 token/s/8x4090
- Serving **efficiency** optimization - model pipelining through network
- **Correctness** & **integrity** - optimistic & deterministic serving

Xiaoyuan Liu, UC Berkeley, 2024