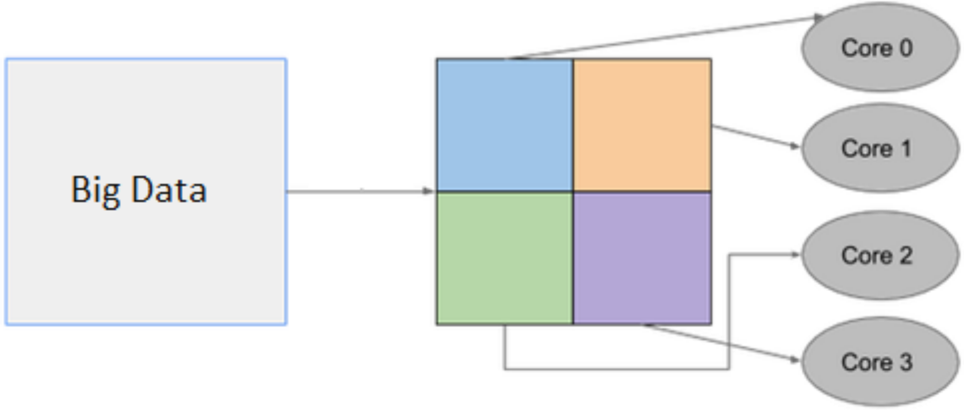
CS-723
EFFICIENTLY SCALING TRANSFORMER INFERENCE

Presented by Bugra Eryilmaz

PARALLEL SYSTEMS ARCHITECTURE LAB

What is partitioning?

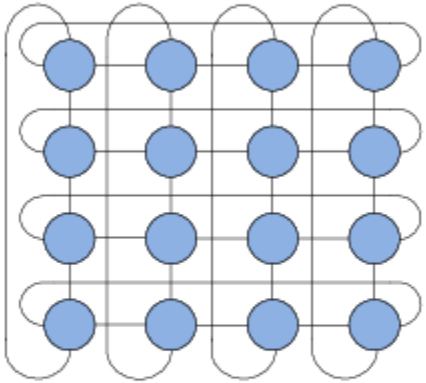
- Multi-device scenario
- Data is too big to fit
- Divide it to multiple devices
- Different partitions are possible



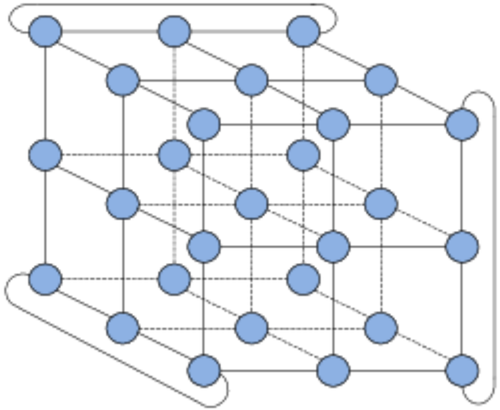
Topology

- Physical layout of chips
- 3D Torus topology is used in the paper
- 3 axis of partitioning

1-D Torus (4-ary 1-cube)



2D Torus (4-ary 2-cube)



3D Torus (3-ary 3-cube)

Q1) What is the problem?

- Large memory footprint
  - PaLM has 540B parameters
  - Single chip cannot store the model
  - Large memory traffic
- Tight latency requirements
- Partitioning scheme effects possible utilization and latency

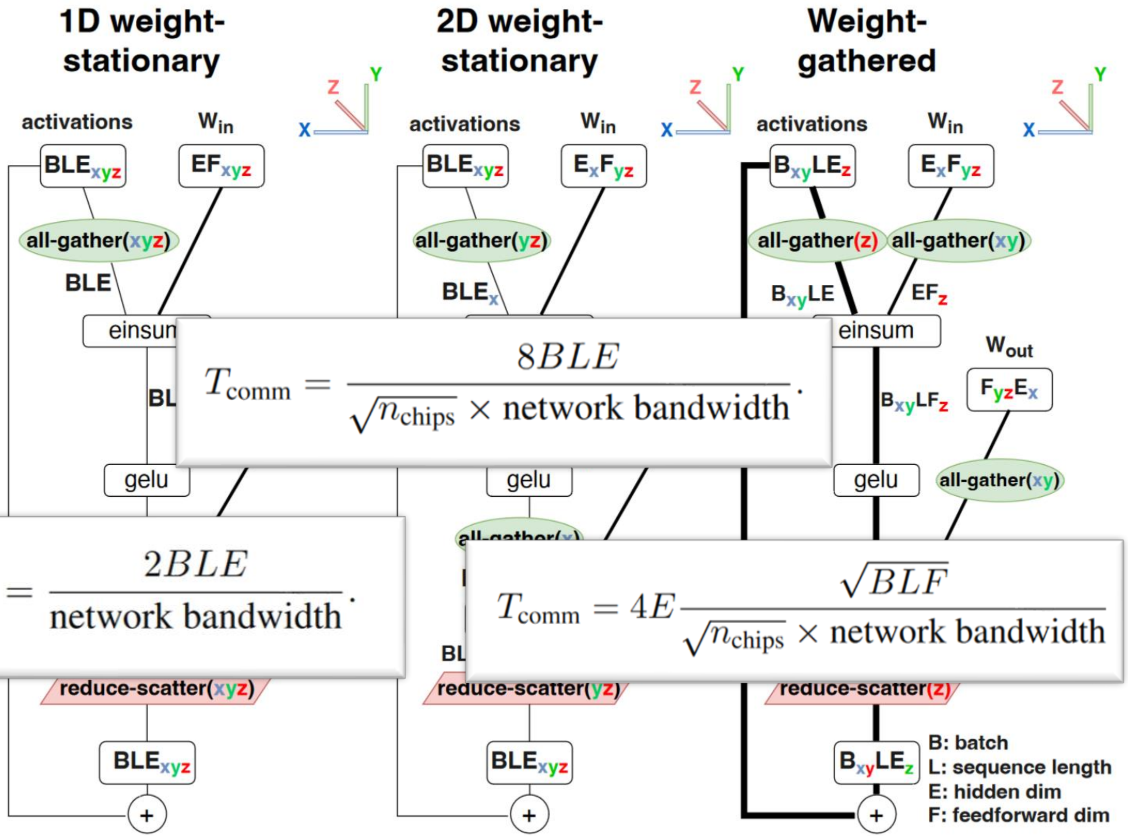
Prior work do not explain the tradeoffs in partitioning schemes!

Q2) What are the insights?

- Memory traffic is the main latency limiting factor
- Chip count, batch size and partition determines the traffic
- Analytical solutions helps understanding the tradeoffs

Q3) What is the solution?

- Identified common cases
- Analytically solved for communication delay



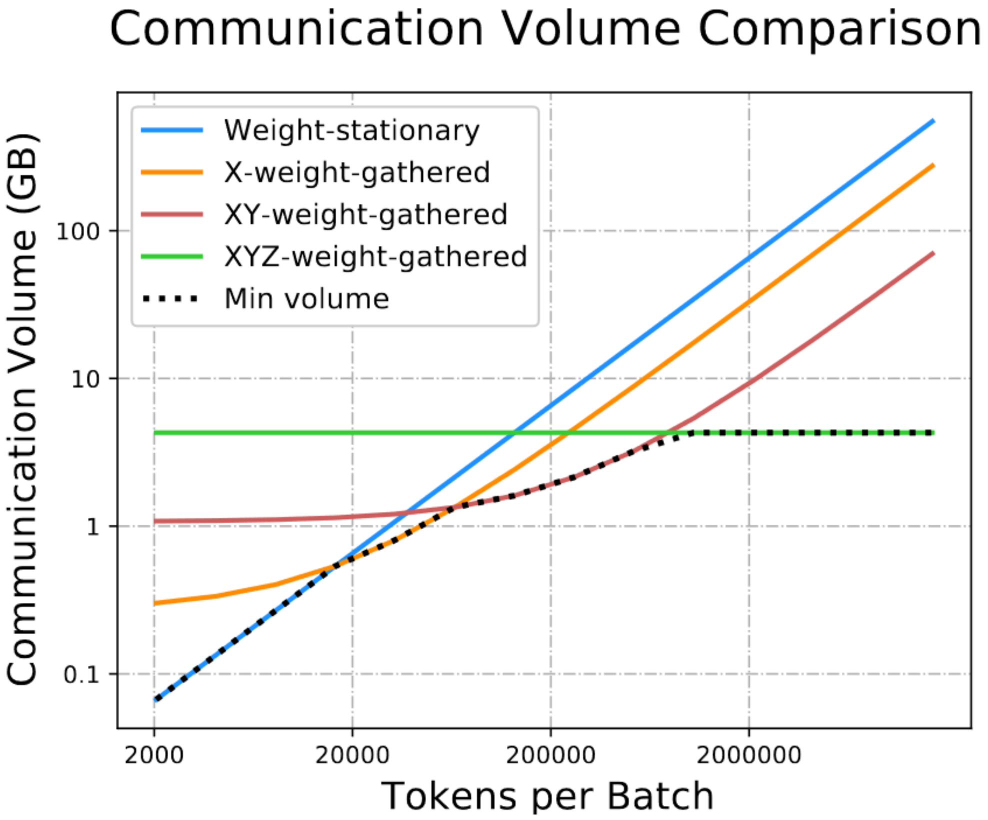
$$T_{\text{comm}} = \frac{2BLE}{\text{network bandwidth}}.$$

$$T_{\text{comm}} = \frac{8BLE}{\sqrt{n_{\text{chips}}} \times \text{network bandwidth}}.$$

$$T_{\text{comm}} = 4E\frac{\sqrt{BLF}}{\sqrt{n_{\text{chips}}} \times \text{network bandwidth}}$$

Q3) What is the solution?

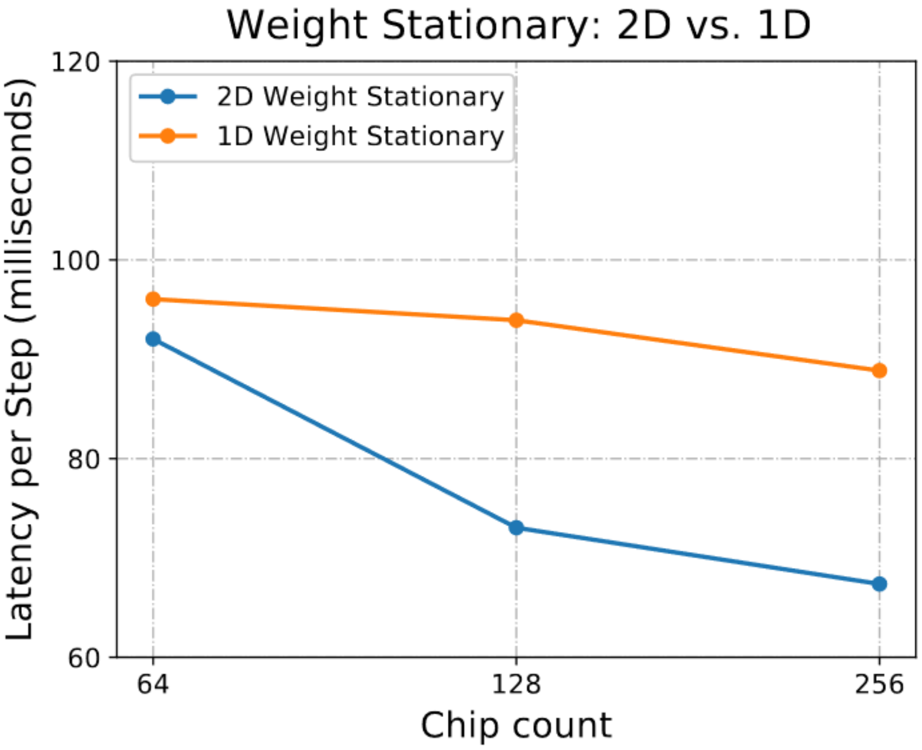
- Identified common cases
- Analytically solved for communication delay
- Showcased the tradeoffs



Q3) What is the solution?

PARALLEL SYSTEMS ARCHITECTURE LAB

- Identified common cases
- Analytically solved for communication delay
- Showcased the tradeoffs



Q4) What is the takeaway message?

- The main bottleneck is memory traffic
- Careful consideration towards partitioning is necessary
  - Partitioning scheme should minimize memory traffic
  - Latency and utilization is the main tradeoff effecting the partition choice

Q5) Will this paper win the test of time?

- No

- The solution is extremely specific
  - Partitioning scheme options can change
  - Optimization dependent
  - Topology dependent

Q6) Why should this paper not have appeared at a top conference?

- Limited explanation on partitioning schemes
- Methodology seems problematic
  - Comparing different hardware
- Paper loses focus a lot

Thank you

CS-723
Orca: A Distributed Serving System for Transformer-Based Generative Models

Presented by Bugra Eryilmaz

PARALLEL SYSTEMS ARCHITECTURE LAB

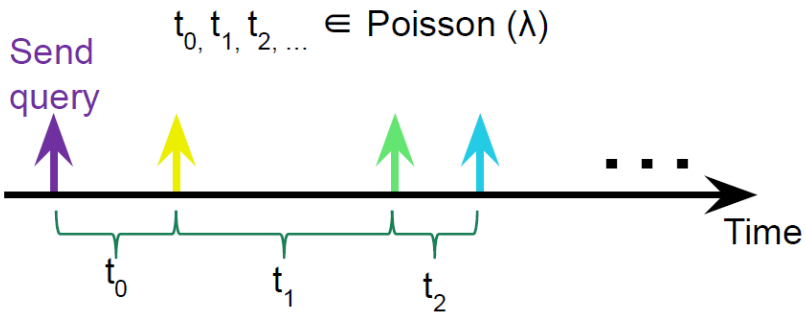
EPFL

Batching

- What is batching?
  - Combining multiple requests
- Why do we need batching?
  - Parameter reuse
  - Utilize the parallelism

Why do we need scheduling?

- Requests arrive randomly
- One resource many consumers
    - E.g., one GPU serving multiple requests
    - Fairness while efficiently utilizing the hardware



Q1) What is the problem?

- Iterative output generation
- Each request runs a different number of iterations
  - Wasted computation for early finishing requests
- Inputs come at different times
  - Queueing delay waiting for the previous batch
- Input shape depends on iteration count and input tokens

Prior work do not address all challenges together!

Q2) What are the insights?

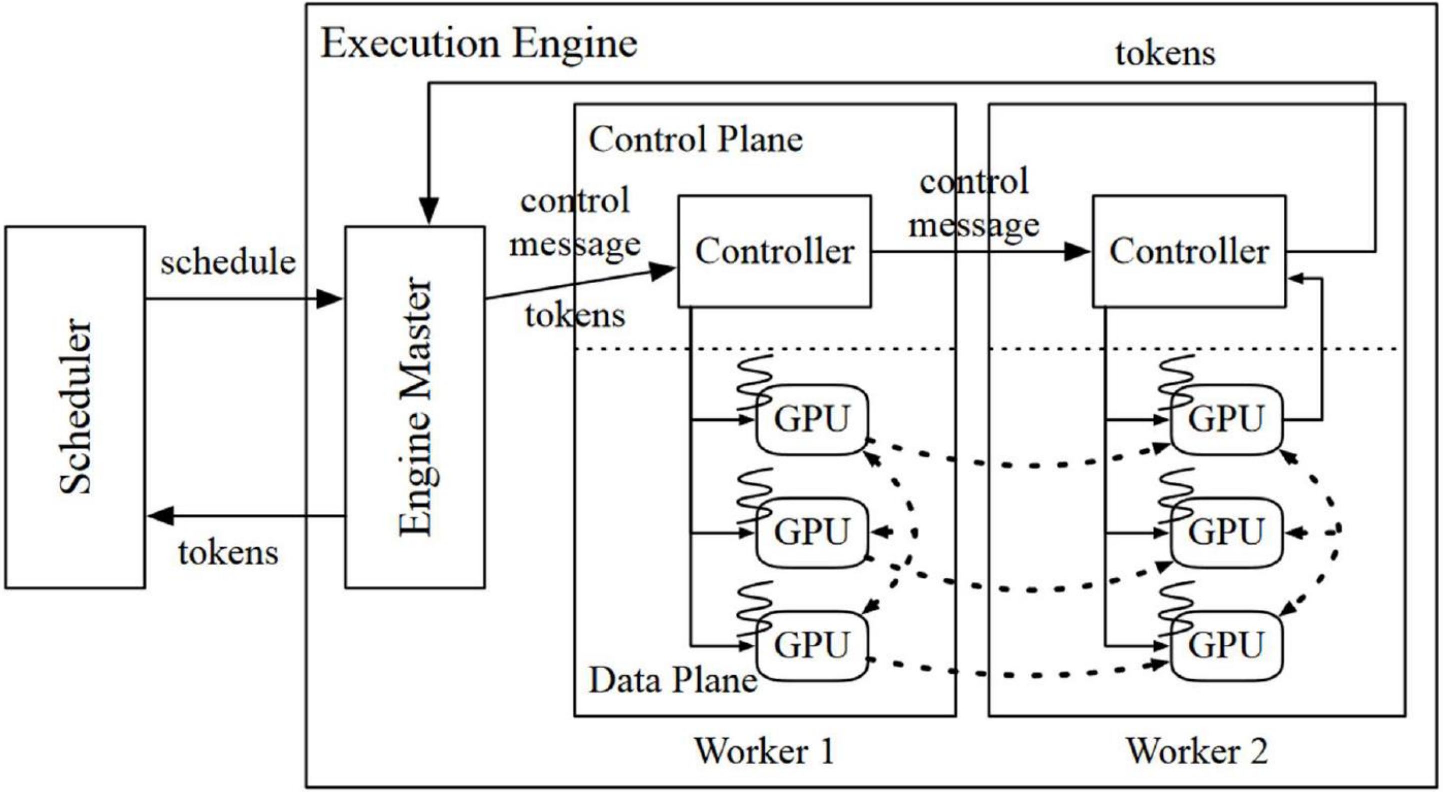
- Scheduling in granularity of iteration
  - Early finishing requests can return
  - Late coming requests can join the in-flight batch

- Attention block does not need to be batched
  - It does not benefit much from batching
  - Input shape is only a problem in attention blocks

Q3) What is the solution?

- Distributed serving system
- Inter and intra layer parallelization
- FIFO based iteration level scheduling
- Selective batching

Q3) What is the solution?



Q4) What is the takeaway message?

- For iterative models we might need iteration level scheduling
  - Early finishing requests can return
  - Late coming requests can join the in-flight batch

Q5) Will this paper win the test of time?

PARALLEL SYSTEMS ARCHITECTURE LAB

- My answer is yes!
- The ideas are applicable to broad areas
  - Only assumption is attention based iterative model
  - Attention and iteration is essential for sequential data
- The solution and ideas are feasible and simple

Q6) Why should this paper not have appeared at a top conference?

- I could not find a problem in the paper
  - Relevant problem
  - Clear and simple insights
  - Simple, feasible and effective solution
  - Fair methodology
  - Maybe it could be in a different venue, but technically fits into OSDI

Thank you