Challenges in Distributed Compute

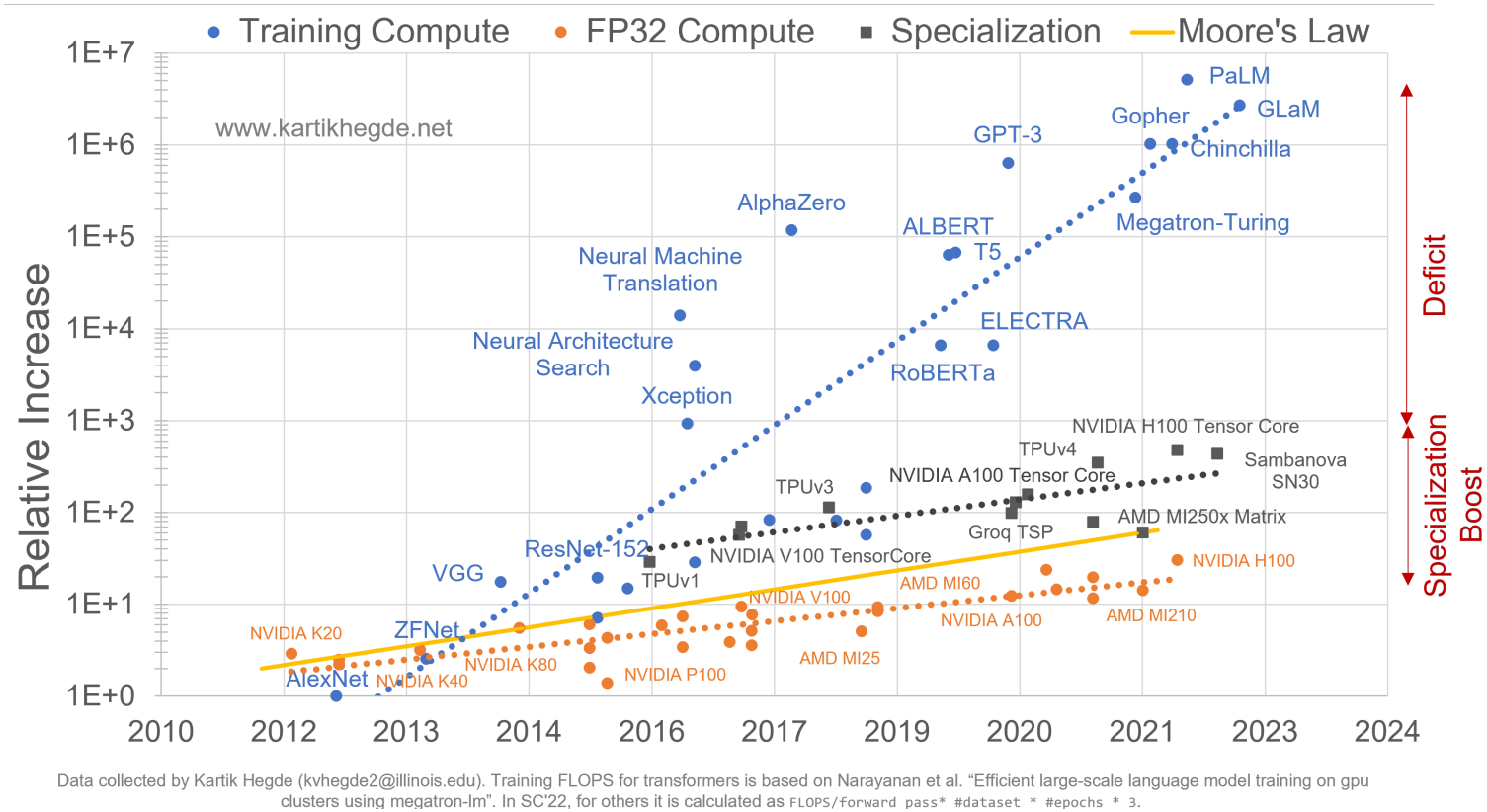
Heterogeneity, contention, and pricing

Parimal Parag

Electrical Communication Engineering
Indian Institute of Science

The growing performance deficit

Deep learning compute demand[1]



Data collected by Kartik Hegde (kvhegde2@illinois.edu). Training FLOPS for transformers is based on Narayanan et al. "Efficient large-scale language model training on gpu clusters using megatron-lm". In SC'22, for others it is calculated as `FLOPS/forward pass* #dataset * #epochs * 3`.

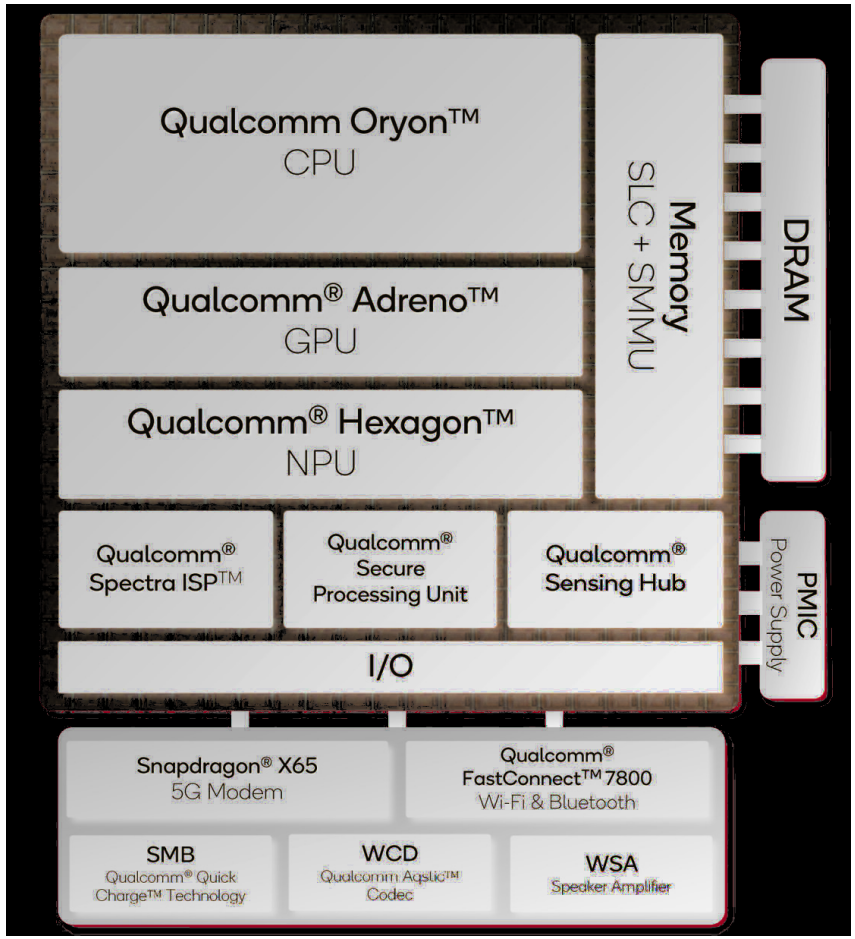
- End of Moore's law and Dennard scaling
- Can no longer keep adding more transistors and increasing core frequencies
- We're accustomed to increasing compute

[1] Image credit: https://kartikhegde.substack.com/p/accelerating-deep-learning-in-the

Heterogeneous Computing[2]

Heterogeneous infrastructure

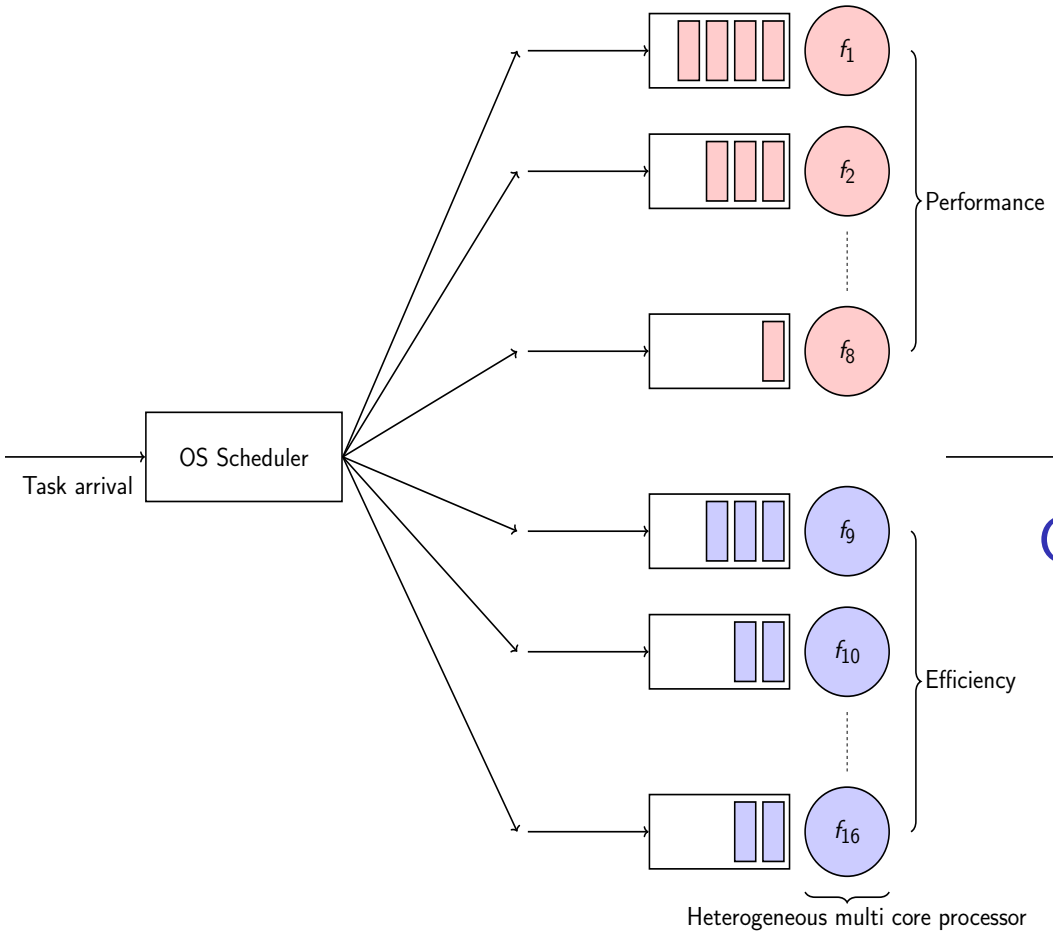
- Different generations of servers and accelerators
- Pooling of all available compute resources (CPUs, GPUs, NPUs)
- Compute resources may be run at different speeds for energy conservation
- Compute cores optimized for different operating regions, to deal with dynamic workloads
- External factors—network bottlenecks, data affinity, etc.



[2] Image: https://www.anandtech.com/show/21445/qualcomm-snapdragon-x-architecture-deep-dive

Energy aware compute scheduling[3]

Heterogeneous multi core system



Objective

- Minimize core energy consumption
- Meet service guarantees such as mean or tail compute latency

Control parameters at cores

- Load balancing
- Operating frequency
- Idle or active state

[3] A. Priya, R. Choudhury, S. Patni, H. Sharma, M. Mohanty, K. Narayanam, U. Devi, P. Moogi, P. Patil, and P. Parag. Energy-minimizing workload splitting and frequency selection for guaranteed performance over heterogeneous cores. *ACM International Conference on Future and Sustainable Energy Systems (e-Energy)*. pp. 308–322, June 2024.

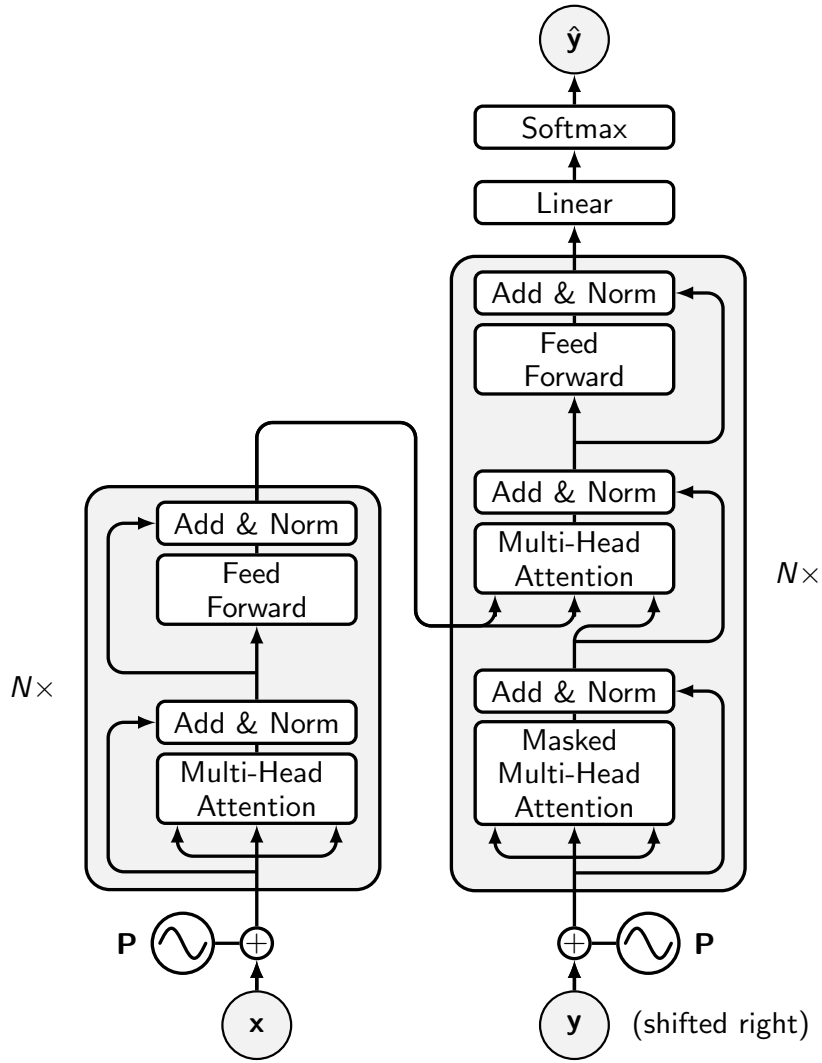
Heterogeneous Computing[4]

Heterogeneous workload

- Training and fine tuning
- Inference

Heterogeneous inference

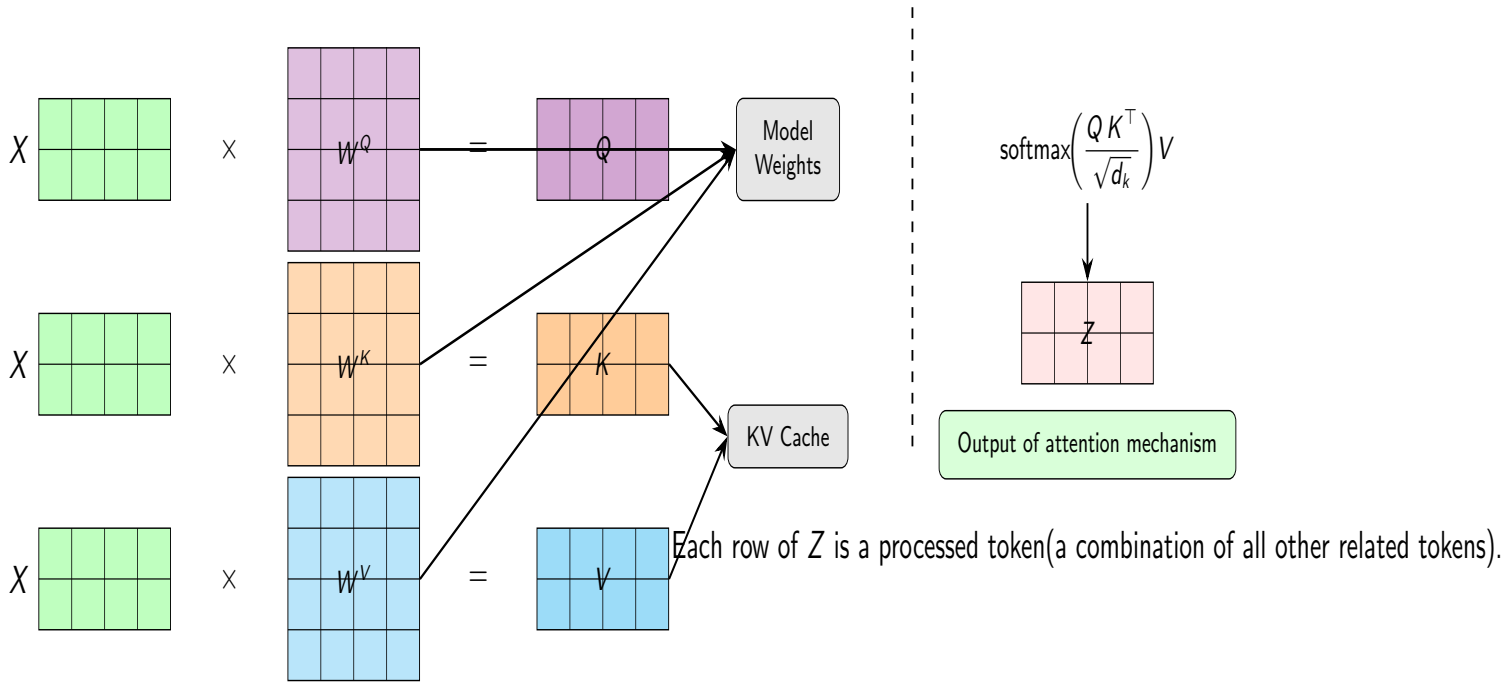
- Conversational AI
- Content generation
- Code completion
- Search or recommendation
- Data analysis & visualization



[4] Image: A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

Inference modeling

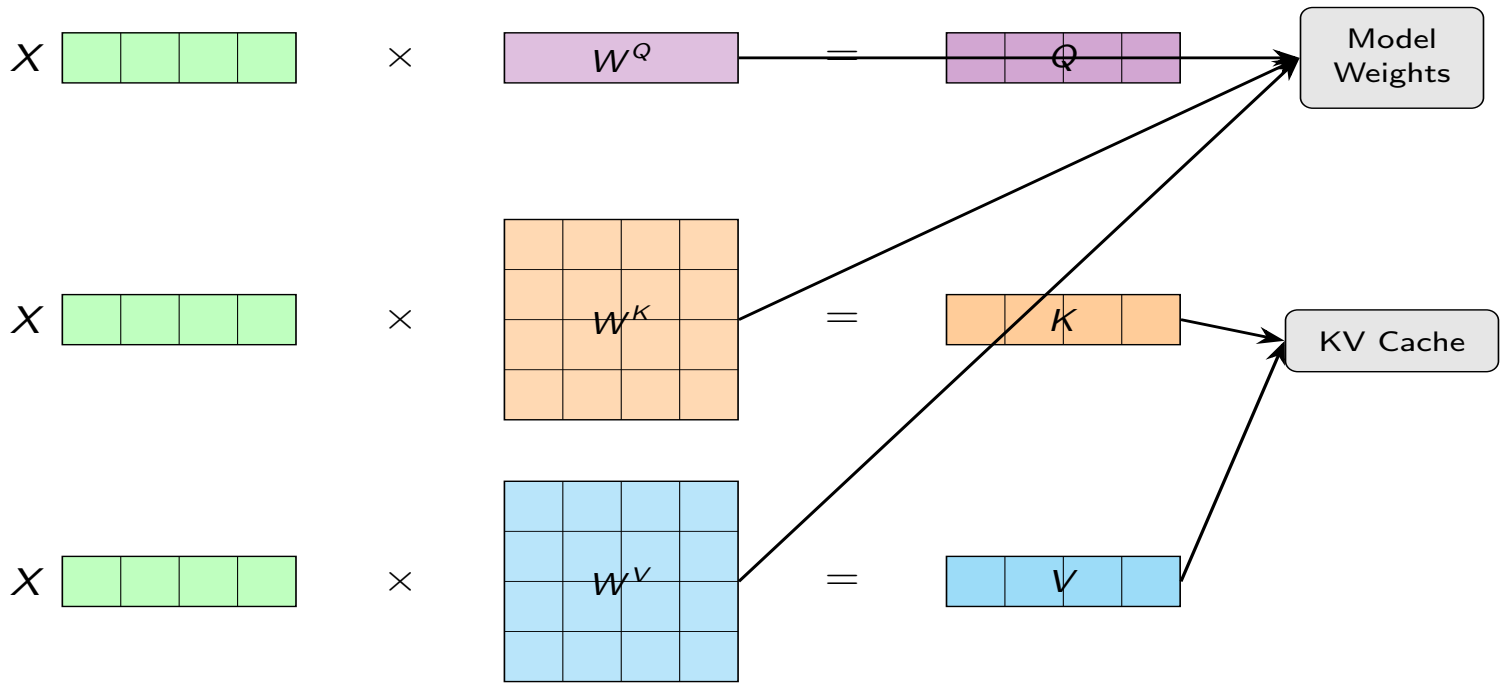
Prefill



- Preprocessing of input tokens to generate first output token
- Computationally challenging task

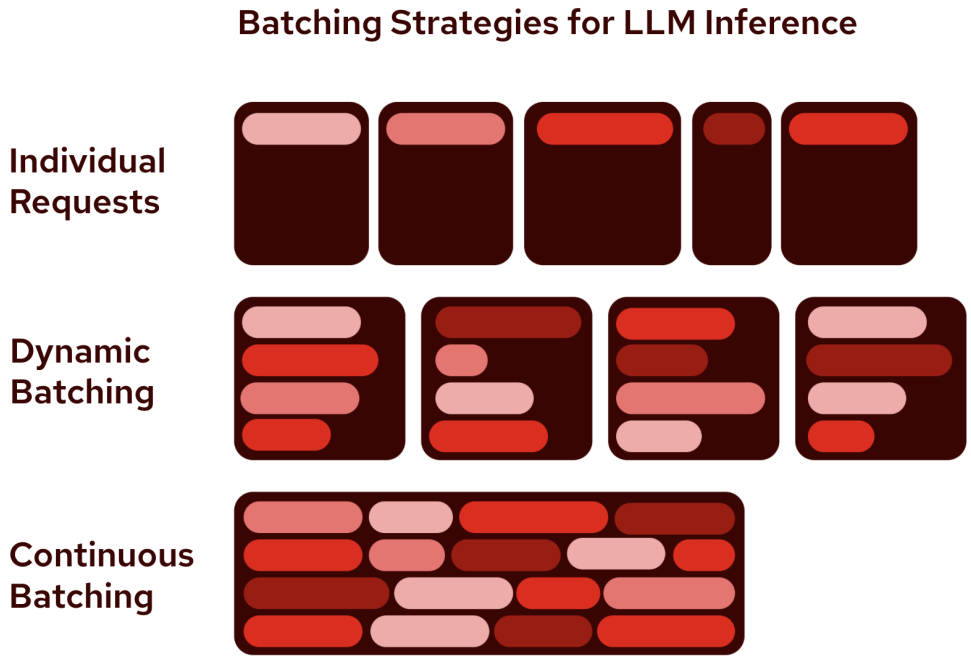
Inference modeling

Decode



- Subsequent generation of output tokens
- No need to repeat previous computations
- Fetching previous computation from memory
- Storing current computation to memory
- Sequential and memory intensive

Speeding up LLMs[6]



- Prefill followed by batched decoding[5]
- Disaggregate prefill and decode
- Simultaneous prefill and decode
- Model parallelism, data parallelism, pipelining
- Heterogeneous combining, quantization, pruning etc.

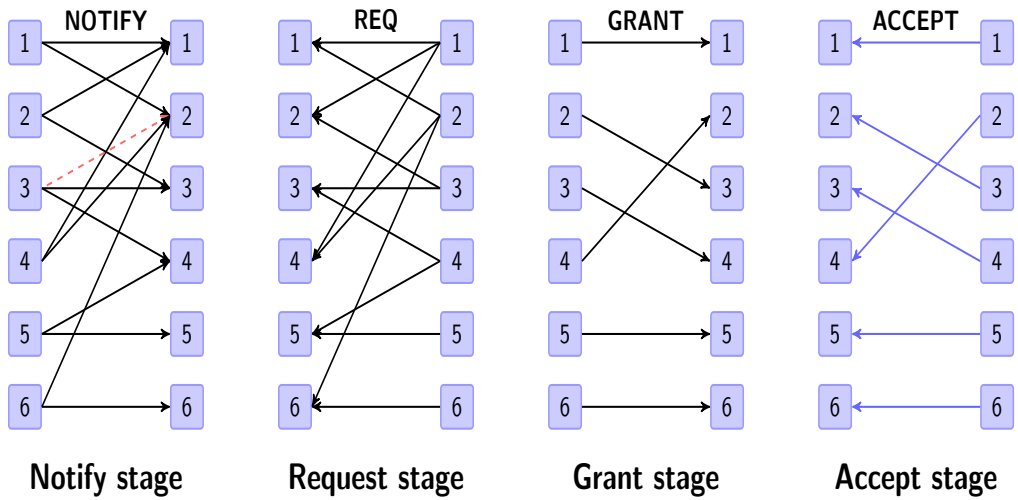
[5] M. Mohanty, G. Bolar, P. Patil, U. Devi, F. George, P. Moogi, and P. Parag. Deferred prefill for throughput maximization in LLM inference. *Workshop on Machine Learning and Systems (EuroMLSys)*, pp. 100–106, Mar 2025.

[6] Image: https://nlpcloud.com/llm-inference-optimization-techniques.html

Communication for distributed compute

Need for new protocols

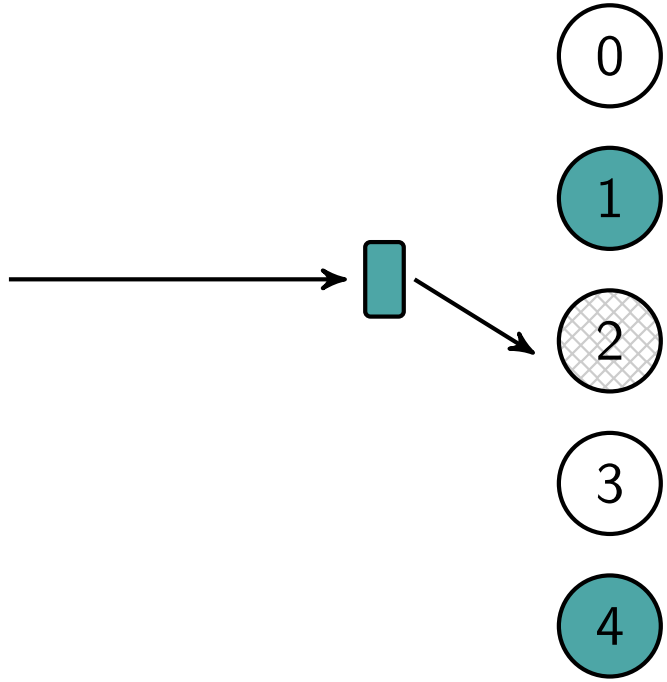
- Fast links, low delay, and low latency desirable
- TCP not a good model for compute messages



- Short messages transmitted without delay, long messages need receiver permission
- Matching at the heart of the load balancing communication messages [7]

[7] M. Mohanty, G. Bolar, P. Patil, A. Ganesh, J. F. Chamberland, and P. Parag. Bipartite matching under communication constraints.

Cluster pricing



Approaches

- Service guarantees for higher paying customers
- Switching between models depending on prices
- Centralized routing and pricing for compute clusters [8]
- Decentralized routing and pricing for compute clusters[9]

[8] A. Krishnan K. S., C. K. Singh, S. T. Maguluri, and P. Parag. Optimal pricing in multi server systems. *Performance Evaluation*. 154, 102282, Apr 2022.

[9] D. Narasimha, S. Nomula, S. Shakkottai, and P. Parag. The Power of Two in Large Service-Marketplaces. *IEEE International Conference on Computer Communications (INFOCOM)*. May 2025.

Acknowledgements





CENTRE FOR
NETWORKED INTELLIGENCE
Indian Institute of Science

IBM

intel

Qualcomm

NRF National Research Foundation

References

- D. Narasimha, S. Nomula, S. Shakkottai, and P. Parag. The Power of Two in Large Service-Marketplaces. *IEEE International Conference on Computer Communications (INFOCOM)*, May 2025.
- M. Mohanty, G. Bolar, P. Patil, U. Devi, F. George, P. Moogi, and P. Parag. Deferred prefill for throughput maximization in LLM inference. *Workshop on Machine Learning and Systems (EuroMLSys)*, pp. 100–106, Mar 2025.
- M. Mohanty, G. Gautam, V. Aggarwal, and P. Parag. Analysis of fork-join scheduling on heterogeneous parallel servers. *IEEE/ACM Transactions on Networking*. 32(6):4798–4809, Dec 2024.
- A. Priya, R. Choudhury, S. Patni, H. Sharma, M. Mohanty, K. Narayanam, U. Devi, P. Moogi, P. Patil, and P. Parag. Energy-minimizing workload splitting and frequency selection for guaranteed performance over heterogeneous cores. *ACM International Conference on Future and Sustainable Energy Systems (e-Energy)*. pp. 308–322, Jun 2024.
- R. Jinan, A. Badita, T. P. Bodas, and P. Parag. Load balancing policies without feedback using timed replicas. *Performance Evaluation*. 162, 102381, Nov 2023.
- R. Jinan, G. Gautam, P. Parag, and V. Aggarwal. Asymptotic analysis of probabilistic scheduling for erasure-coded heterogeneous systems. *ACM SIGMETRICS Performance Evaluation Review*. 50(4):8–10, Mar 2023.
- A. Krishnan K. S., C. K. Singh, S. T. Maguluri, and P. Parag. Optimal pricing in multi server systems. *Performance Evaluation*. 154, 102282, Apr 2022.
- A. Badita, P. Parag, and J.-F. Chamberland. Latency analysis for distributed coded storage systems. IEEE Transactions on Information Theory. 65(8):4683–4698, Aug 2019.