Bipartite matching under communication constraints

Moonmoon Mohanty, Gautham Bolar, Preetam Patil, Parimal Parag
Ayalvadi J. Ganesh, Jean-Francois Chamberland

Communication, learning, information, and processing seminar, Aug 05, 2026

Anusandhan National Research Foundation

CENTRE FOR NETWORKED INTELLIGENCE
Indian Institute of Science

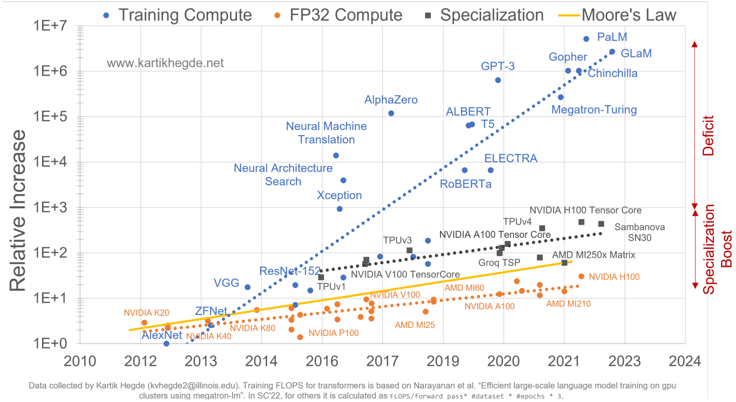
IBM

Qualcomm

intel

The growing performance deficit

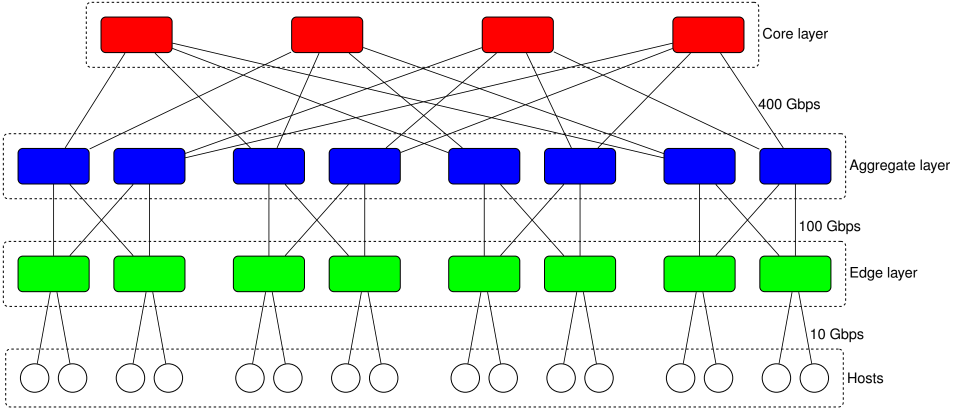
Deep learning compute demand[1]



- 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 achievable only by scaling out through data center network (DCN)

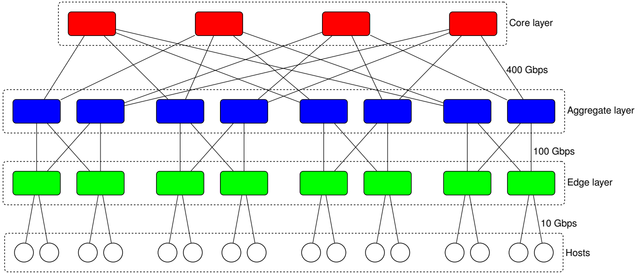
[1] Image credit: https://kartikhegde.substack.com/p/accelerating-deep-learning-in-the

DCN for compute



- DCN traffic comprises of
  - Latency sensitive compute applications (short messages)
  - Bandwidth hungry applications (long messages)
- The link from host to edge layer switch is the bottleneck
- End-hosts can be senders and receivers

Why not use TCP for networked compute?



Inadequacies of TCP for networked compute

- Not all messages are equal (TCP treats them equally)
- TCP congestion control operates at RTT (scale of microseconds in DCN)
  - Too high a latency for networked compute
  - High DCN bandwidth leads to large queue buildup and latency

Transport protocols for networked compute

- Scheduling algorithms for input-queued switches in high-speed LANs and WAN routers
  - Parallel iterative matching (PIM)[2]
  - Iterative round robin scheduling algorithm (iSLIP),[3] an improvement over PIM
- Transport protocols for networked compute
  - Priority based receiver driven Homa[4] improvement over pHost, NDP
  - Data center parallel iterative matching (dcPIM)[5]

---

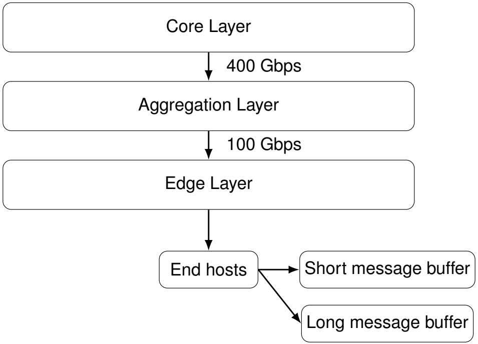
[2] Thomas E Anderson, Susan S Owicki, James B Saxe, and Charles P Thacker. High-speed switch scheduling for local-area networks. *ACM Transactions on Computer Systems, 11(4):319–352, Nov. 1993.*

[3] Nick McKeown. The iSLIP Scheduling Algorithm for Input-queued Switches. *IEEE/ACM Transactions on Networking, 7(2):188–201, Apr. 1999.*

[4] Behnam Montazeri, Yilong Li, Mohammad Alizadeh, and John Ousterhout. 2018. Homa: A Receiver-driven Low-latency Transport Protocol Using Network Priorities. *ACM SIGCOMM 2018.*

[5] Qizhe Cai, Mina Tahmasbi Arashloo, and Rachit Agarwal. dcPIM: Near-optimal proactive datacenter transport. *ACM SIGCOMM 2022.*

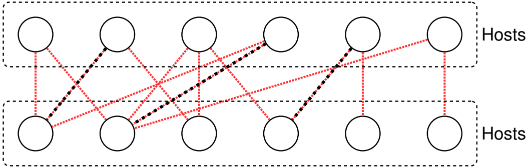
Transport protocol for networked compute



Proposals for transport protocols for compute

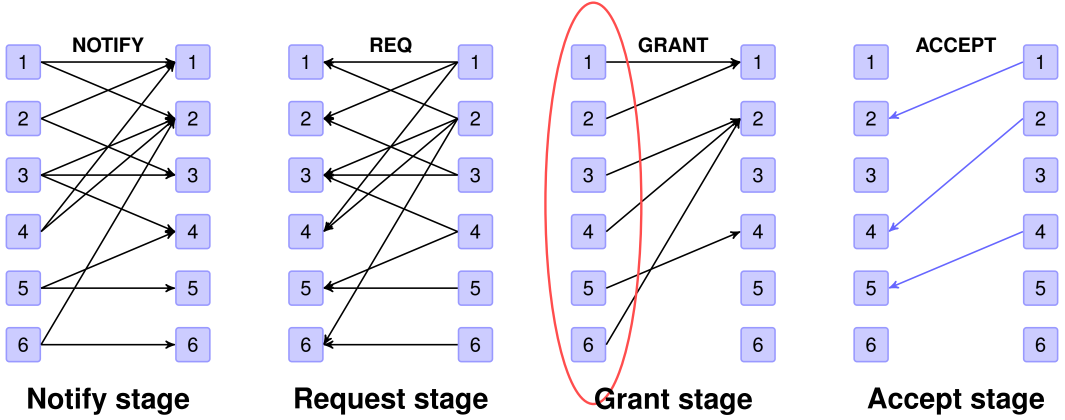
- Distinguish among control, short compute messages, and long persistent messages
- Short messages are prioritized and no congestion control for them
- Congestion control for long messages via receiver grant mechanism
- Matching of sender receiver pairs for efficient utilization of bottleneck edge links

Modeling assumptions



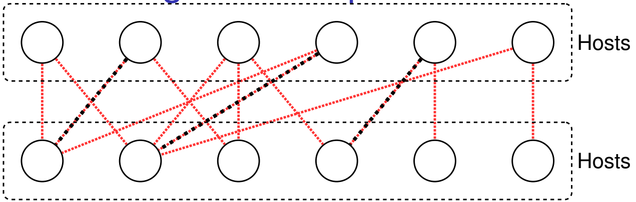
- Time is divided into discrete slots in order of microseconds
- Each slot consists of a matching and data transmission phase
- Matching phase
  - forming connection graph based on senders with long messages for receivers
  - collecting neighborhood information
  - matching algorithm
- Connection graph is modeled as a D-out random graph
  - $D = \mathrm{Bin}(N, d/N)$ yields Erdős-Rényi graph
  - $D = d$ yields deterministic out degree
- Data transmission phase: parallel long message transmission for each matched pair
- Goal: Establish a large bipartite matching in each time slot

Control message exchange during matching phase



- Bipartite matching sender receiver pairs is at the heart of the problem
- Each round consists of matching and data transmission phases
- Distributed matching with local information to keep the matching phase small

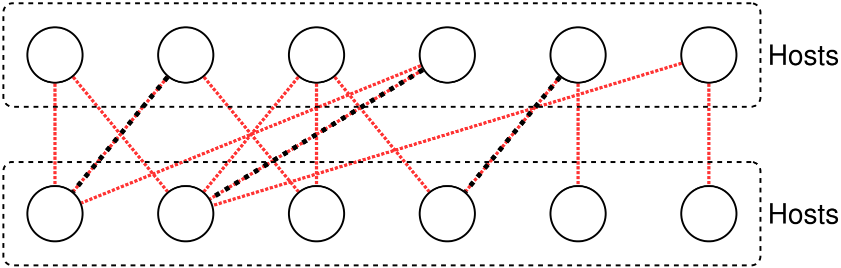
Matching algorithms using Grant requests



Related works

- Centralized bipartite matching
  - Hopcroft-Karp: computes maximum matchings in polynomial time
  - Hungarian: polynomial-time algorithm widely used in applications like client selection or resource allocation to maximize total weight
- Distributed matching
  - Random bipartite matching
    - Uses one neighborhood information
    - Picks one of its intended receivers uniformly at random
  - Proposal: Degree-Biased (DB) (α) random bipartite matching
    - Uses two neighborhood information
    - Biased selection of one of its receivers

Degree-biased selection

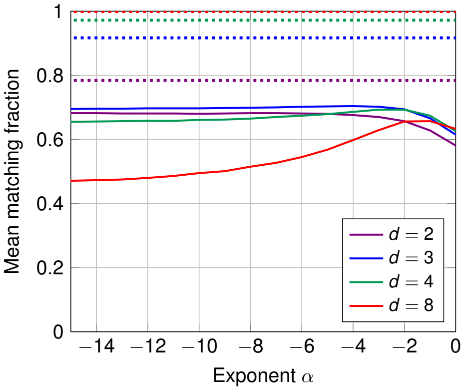


DB(α) algorithm

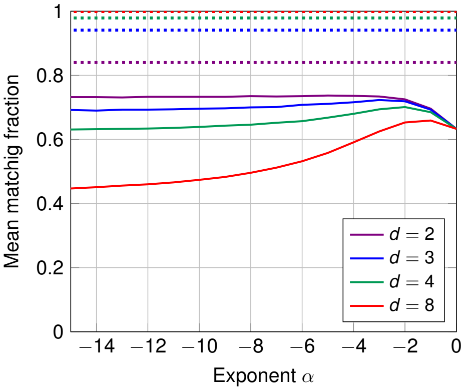
- Edge $(u, v) \in E$ selection probability $\propto \deg_v^{\alpha}$
- $\alpha = 0$: uniform selection
- $\alpha < 0$: prefer low degree neighbors
- $\alpha = -\infty$: greedy selection

Mean matching fraction

Impact of changing exponent α



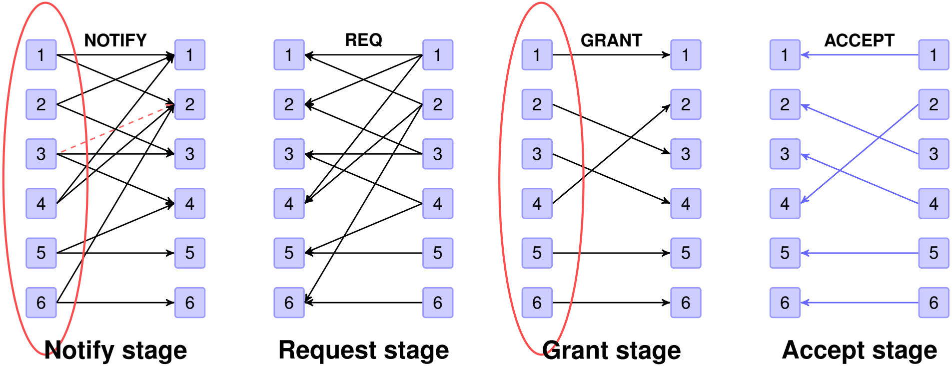
(a) $D \sim \mathrm{Bin}(N, d/N)$



(b) Deterministic $D = d$

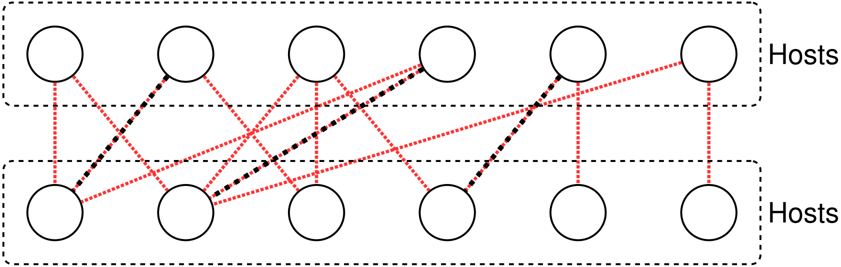
- Higher mean sender degree can counterintuitively reduce matching performance

2CGS algorithm for notification and grant selection



- Random thinning of one neighborhood in the notify stage
- Uses two neighborhood information built during the request stage
- Greedy selection of a minimum degree neighbor uniformly at random in grant stage

Random thinning of connection graph

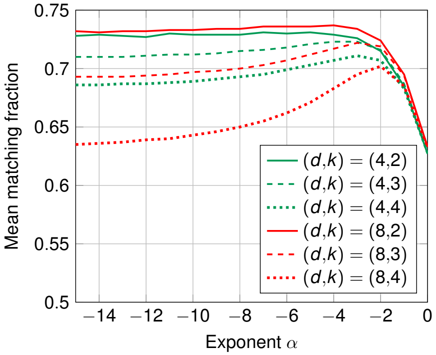


Thinned connection graph

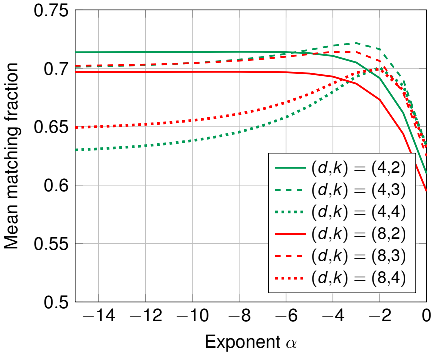
- $\max(k)$ thinning yields a $(D \wedge k)$-out random graph
  - Deterministic d-out thins to k-out
  - $\mathrm{Bin}(N, d/N)$-out thins to $\mathrm{Bin}(N, d/N) \wedge k$-out
- *i.i.d.* thinning with probability k/d leads to $\mathrm{Bin}(D, k/d)$-out random graph
  - Deterministic d-out thins to $\mathrm{Bin}(d, k/d)$-out
  - $\mathrm{Bin}(N, d/N)$-out thins to $\mathrm{Bin}(N, k/N)$-out

Mean matching fraction

Impact of thinning



(a) $D \sim \mathrm{Bin}(N, d/N)$ with $\max(k)$

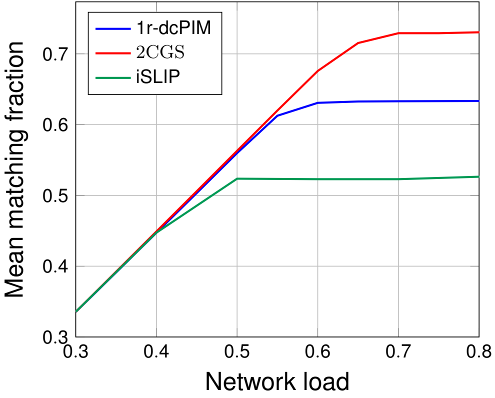


(b) Deterministic $D = d$ with $\mathrm{Bern}(k/d)$

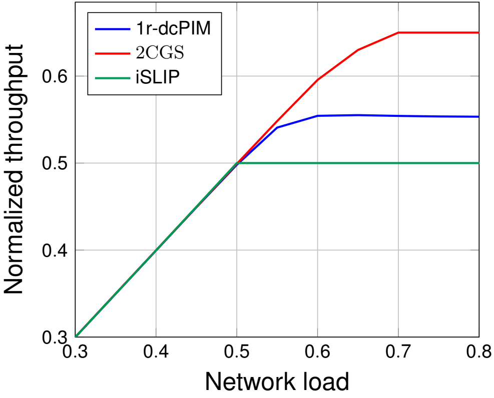
- Thinning to low out-degree improves decentralized matching performance
- $\max(2)$ yields the strongest results

Experimental results

Performance comparison with 1r-dcPIM, and iiSLIP



(a) Mean matching fraction

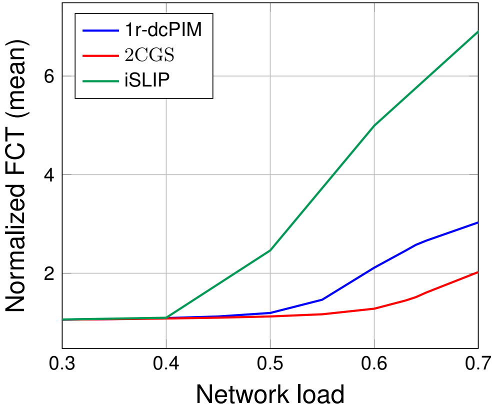


(b) Normalized throughput

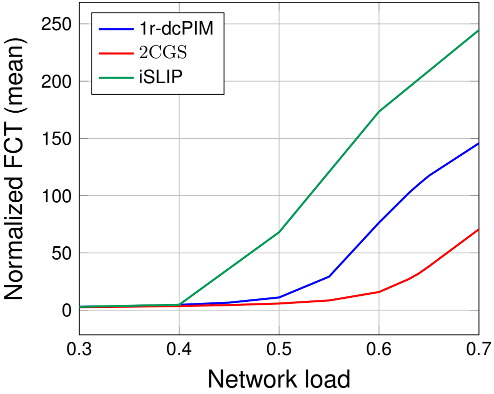
Experimental results

Performance comparison with 1r-dcPIM, and iiSLIP

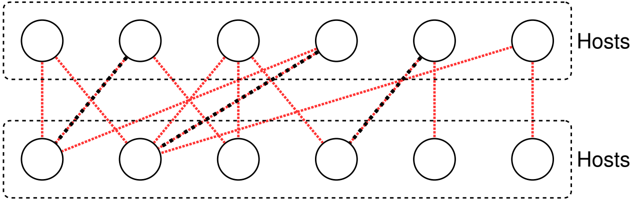
Flow completion time



(a) Short message FCT



(b) Long message FCT



Mean matching fraction

- Probability of a receiver being connected
- No connection to a receiver $=$ having no edges from senders
- *i.i.d.* receiver selection probability across all senders

Uniform selection

- Probability of a receiver being picked by any sender is $P\{D > 0\}/N$
- Mean matching fraction $1 - \left(1 - \frac{P\{D>0\}}{N}\right)^N \approx 1 - e^{-P\{D>0\}}$

Mean matching fraction

- Probability of a receiver being connected
- No connection to a receiver $=$ having no edges from neighboring senders
- Neighboring senders have a size biased distribution identical to D^*
- Under some technical conditions, neighboring senders of a receiver have independent size biased degree distributions

Greedy selection for a receiver

- Each neighboring sender picks the receiver with probability $\frac{1}{X+1}$ if
  - All neighboring receivers have higher or equal degrees
  - X neighboring receivers have identical degree
- receiver selection probability by a neighboring sender

Useful identities

Degree distribution for receivers: $\mathbb{E}D = d$

- Probability for an edge to be present d/N
- Receiver degree $\mathrm{Bin}(N, d/N)$ irrespective of distribution D

Size biased distribution

- Degree D^* distribution of connected sender is $(k/d)P\{D = k\}$ and

$$\mathbb{E}\left[\frac{z^{D^*}}{D^*}\right] = \frac{G_D(z) - P\{D = 0\}}{d}$$

Binomial identity

- For $X \sim \mathrm{Bin}(N, p)$, we have

$$\mathbb{E}\left[\frac{\theta^{X+1}}{X+1}\right] = \frac{(1-p+p\theta)^{N+1} - (1-p)^{N+1}}{(N+1)p}$$

Greedy selection for a receiver with s senders

- Let q, Q be pmf and cdf for binomial distribution
- Each neighboring sender picks the receiver with probability $\frac{1}{X+1}$ if
  - All neighboring receivers have degrees $\geqslant s-1$, this event probability is $(1 - Q_{s-2}(N-1, d/N))^{D^*-1}$
  - Conditioned on the previous event, X neighboring receivers have identical degree where

$$X \sim \text{Bin}\Big(D^* - 1, \underbrace{\frac{q_{s-1}(N-1, d/N)}{1 - Q_{s-2}(N-1, d/N)}}_{\text{probability } p}\Big)$$

- Conditional receiver selection probability by a neighboring sender $\frac{1-(1-p)^{D^*}}{D^* p}$
- Receiver selection probability by a neighboring sender is

$$\mathbb{1}_{\{u \sim v\}} \frac{\bar{Q}_{s-2}(N-1, \frac{d}{N})^{D^*} - (\bar{Q}_{s-1}(N-1, \frac{d}{N}))^{D^*}}{D^* q_{s-1}(N-1, \frac{d}{N})}$$

Mean matching fraction

Greedy selection

- Conditional probability of an edge being selected when neighborhood of sender is known and receiver degree is s and q, Q are pmf and cdf of binomial distributions

$$\mathbb{1}_{\{u\sim v\}}\frac{\bar{Q}_{s-2}(N-1,\frac{d}{N})^{D^*}-(\bar{Q}_{s-1}(N-1,\frac{d}{N}))^{D^*}}{D^* q_{s-1}(N-1,\frac{d}{N})}$$

- Under some technical conditions, neighbors of a receiver have independent size biased degree distributions
- Limiting mean matching fraction is lower bounded by

$$1-\sum_{s\geqslant 0}e^{-d}\frac{d^s}{s!}(1-f(s))^s$$

where π and Π are pmf and cdf for Poisson random variables with parameter $\mathbb{E}D$ and

$$f(s)\triangleq\frac{G_D(\bar{\Pi}_{s-2})-G_D(\bar{\Pi}_{s-1})}{d\pi_{s-1}}.$$

Conclusion

Feature of the proposed distributed 2CGS matching algorithm

- Based on two neighborhood information
- Theoretical analysis for greedy selection and D-out random graphs in both sparse and dense regimes
- Improves the mean matching size and hence throughput for long messages
- Reduces control message overheads, improving throughput for short messages as well

Reference

M. Mohanty, G. Bolar, P. Patil, A. J. Ganesh, J. F. Chamberland, P. Parag, "Bipartite matching under communication constraints", *arXiv:2604.10744*, Jun 2026.