通信帯域の変動とノード離脱を考慮した分散LLM推論における動的レイヤ再配置の検討

田口 開智[1,a)]　河野 尊[1,b)]　神野 敦哉[1,c)]　那須 渚[1,d)]　小比賀 亮仁[1,e)]　小板 隆浩[1,f)]

1. はじめに

大規模言語モデル（LLM）はパラメータ数が大きく，単一の計算機では推論できない場合がある．そこで，モデルを分割し，複数ノードで協調して推論する分散推論が研究されている [1]．ノード間の通信帯域は時間とともに変動し，一時的に大きく低下することもある [2]．この変動は推論性能を下げ，推論遅延を最小とする配置も変えうる．

環境変化に対応する研究として，Petals は担当ブロックを定期的に見直し，スループットの改善が 20%の固定しきい値を超える場合に限り変更するが，見直しに用いる性能は参加時の測定値である [1]．SpotServe は移行コストを最小化しつつ並列化構成を変更するが，同種 GPU と中断の事前通知を前提とする [3]．Petals のような固定しきい値は再配置にかかる時間や低下の継続時間を考慮せず，帯域低下に用いるとすぐ回復する低下でも再配置しうる．そこで本稿では，これらを考慮して再配置の可否を判断する枠組みを検討し，帯域低下時に再配置が有効な条件を考察する．

2. 研究課題

本研究では，性能の異なる複数ノードが LLM の連続するレイヤ区間を分担し，単一リクエストを逐次生成する環境を考える．帯域低下では現在の配置のまま継続できるため，再配置の可否が課題となる．本研究では判断の要因として，再配置のコスト，低下の継続時間，遅延の改善幅に着目する．再配置中は推論が停止し，低下した回線で重みを転送すればその時間も延びる．低下は一時的な場合があり，早く回復すれば利得は小さい [2]．また，デコード時にノード間で送る中間データは小さく，帯域低下による遅延の増加は小さい可能性がある．一方，ノード離脱時は再配置が必須であり，残りの生成時間が最小の候補を選ぶ．

[1] 同志社大学
[a)] ctwp0168@mail4.doshisha.ac.jp
[b)] cgij1036@mail4.doshisha.ac.jp
[c)] ctwp0126@mail4.doshisha.ac.jp
[d)] ctwn0144@mail4.doshisha.ac.jp
[e)] akohiga@mail.doshisha.ac.jp
[f)] tkoita@mail.doshisha.ac.jp

3. 検討方針

1 トークンの生成時間を遅延とし，各ノードの計算時間と，各リンクの伝搬遅延および中間データのサイズを帯域で割った転送時間の和で表す．再配置コスト C は再配置開始から推論再開までの停止時間で，重みの取得，再初期化，KV キャッシュの移送・再計算を含む．120B モデルの読込みに 2 分以上を要した例もある [3]．再配置の候補は，メモリ制約内の境界移動，担当ノード交換，区間統合で作る．

帯域低下時の再配置の可否は，再配置する場合としない場合で，残りの生成を終えるまでの時間を比べて決める．低下中の 1 トークンあたりの遅延を現在の配置で L_{cur}，再配置後で L_{new} とし，L_{new} と C は，事前に測った各ノードの計算時間と測定した帯域から候補ごとに見積もる．判断時点からの低下の継続時間を T とし，初期評価では既知とする．$L_{\mathrm{new}} < L_{\mathrm{cur}}$ の候補で，$T > C$，生成が回復後も続き，回復後の遅延が等しいと仮定すると，両者の差は低下中に生成できるトークン数 T/L_{cur} と $(T-C)/L_{\mathrm{new}}$ の違いで決まり，条件は $T > C\,L_{\mathrm{cur}}/(L_{\mathrm{cur}} - L_{\mathrm{new}})$ となる．回復後に現在の配置が速ければ条件はより厳しい．

4. まとめと今後の展望

この条件は，単純化した仮定の下では，再配置が有効なのは低下が C より十分長く，遅延の改善幅が大きい場合に限られることを示す．よって C と T を見積もる判断の検討は妥当と考える．今後はシミュレーションで，再配置しない場合や固定しきい値と推論完了時間・再配置回数を比べて検証し，T の推定誤差や離脱時の配置選択も調べる．

参考文献

[1] Borzunov, A. et al.: Distributed Inference and Fine-tuning of Large Language Models Over The Internet, *Proc. NeurIPS*, pp. 12312–12331 (2023).
[2] Uta, A. et al.: Is Big Data Performance Reproducible in Modern Cloud Networks?, *Proc. USENIX NSDI*, pp. 513–528 (2020).
[3] Miao, X. et al.: SpotServe: Serving Generative Large Language Models on Preemptible Instances, *Proc. ASPLOS*, pp. 1112–1127 (2024).

© 2018 Information Processing Society of Japan