NuPS: A Parameter Server for Machine Learning with Non-Uniform Parameter Access
SIGMOD 2022

Alexander Renz-Wieland, Rainer Gemulla, Zoi Kaoudi, Volker Markl

TU Berlin, Universität Mannheim, BIFOLD

June 2022
Research Area

- Distributed machine learning
- Parameter servers
The Problem: Communication Overhead

Existing Parameter Servers are slower than the single node!

ComplEx 500 on Wikidata5M (SGD with AdaGrad)
Similar results for Word2Vec and Matrix factorization
Non-Uniform Parameter Access

- Parameter servers do the same thing for all parameters
- But different parameters have different access patterns

Sources of non-uniformity: (1) Skew **Skew** and (2) Sampling **Sampling**
Skew in Parameter Servers

Parameter servers use one technique for all parameters

for hot spots  for long-tail parameters

Replication  +  -
Relocation  -  +

... but no technique fits all access patterns
Can we combine techniques and use a well-suited technique for each parameter?
NuPS: a Non-Uniform Parameter Server

Replication + relocation

![Graph showing number of accesses vs parameters]

- Replicate a few hot spot parameters
- Relocate all others
Sampling

- Randomly sampled access

- Sampling schemes: trade off quality for speed
  - Example: local sampling
Sampling Support in Parameter Servers

```python
keys = sample(N)
values = pull(keys)

handle = prepareSample(distr, N)
...
keys, values = pullSample(handle)
```

▶ PS employs sampling schemes *transparently*
Sampling Conformity Levels

- Sampling conformity levels allow for a controlled trade-off between quality and speed:

  (L1) CONFORM

  (L2) BOUNDED

  (L3) LONG-TERM

  (L4) NON-CONFORM

Local sampling

faster

better sampling quality
Experiments

- Three ML tasks:
  - Knowledge graph embeddings
  - Word vectors
  - Matrix factorization

- 8 nodes x 8 threads
NuPS was 6x faster than state-of-the-art parameter servers (PSs) and 6.9x faster than the single node baseline.
Conclusion

- Non-uniform parameter access
  - problem for existing PSs
  - opportunity for NuPS

- NuPS combines multiple management techniques
  - replication and relocation
  - use a suitable one for each parameter

- NuPS supports sampling directly
  - sampling primitive
  - sampling conformity levels

*NuPS: A Parameter Server for Machine Learning with Non-Uniform Parameter Access*
Alexander Renz-Wieland, Rainer Gemulla, Zoi Kaoudi, Volker Markl
Open Issues and Future Work

- Performance knob: which technique to use for a parameter? (heuristic is not optimal)
- Performance knob: when to relocate?
- Sporadic relocation conflicts limit performance

We address these issues in follow-up work:

**Good Intentions:**
Adaptive Parameter Servers via Intent Signaling
Renz-Wieland, Kieslinger, Gericke, Gemulla, Kaoudi, Markl.
https://arxiv.org/abs/2206.00470