Managed Communication and Consistency for Fast, Iterative Data Analytics

Jinliang Wei

Wei Dai, Aurick Qiao,

Qirong Ho (Institute for Infocomm Research, A*STAR),

Henggang Cui, Greg Ganger, Phil Gibbons (Intel Labs),
Garth Gibson and Eric Xing

PARALLEL DATA LABORATORY

Carnegie Mellon University

Managed Communication for ML Training

- Distributed ML training: often bursty communication
  - Communicate only when necessary for consistency
  - Spare network bandwidth exists, sometimes a lot
- Fresher reads improve algorithm performance
  - Use the spare bandwidth but no more and use it well
- Our solution:

  System framework that manages communication
  - Automatically improves algorithm perf w/ spare bandwidth
  - 2-3X speedup

Example: The Netflix Problem

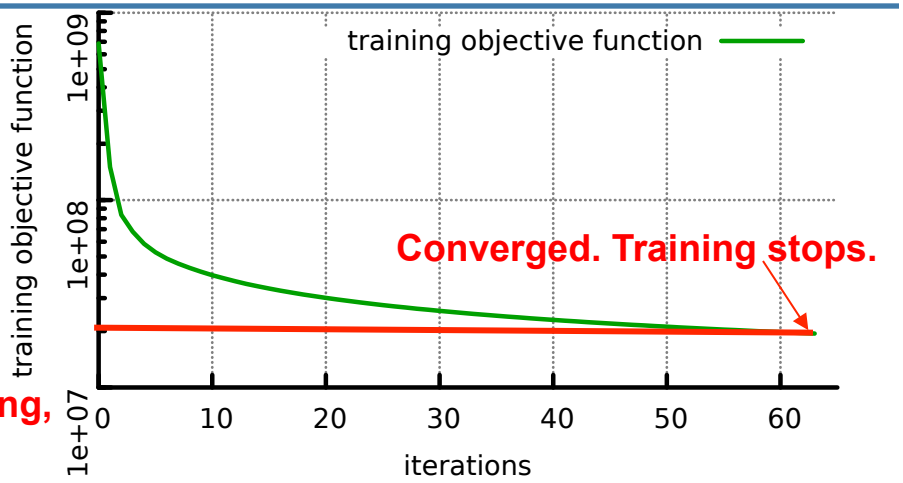
Which movies would my users like to watch?

Collaborative Filtering via ML

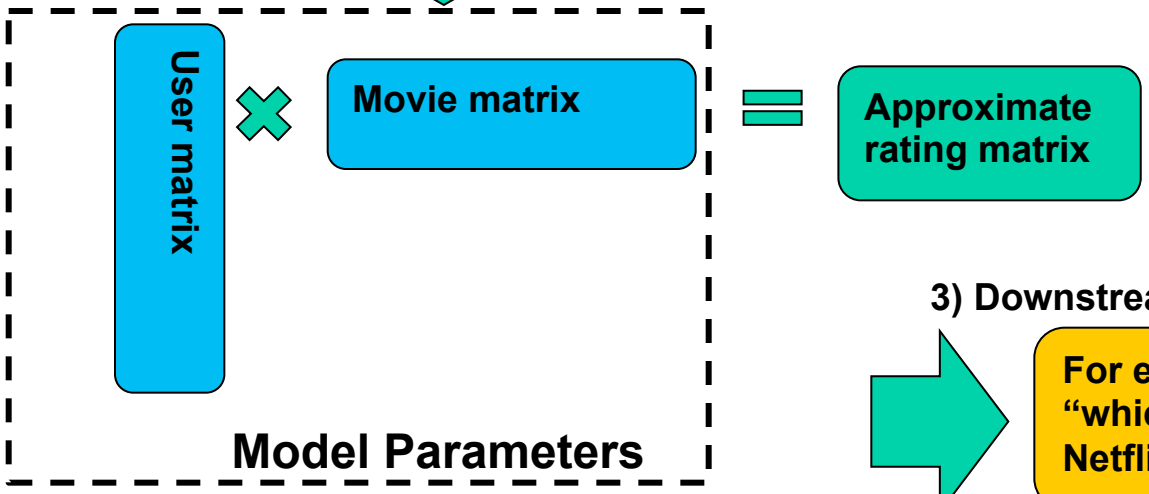
1) Collect training data

	movies			
users	M-A	M-B	M-C	…
Jim	3	?	5	
Bob	2	5	?	
…				

Training data



2) Batch training, often iterative



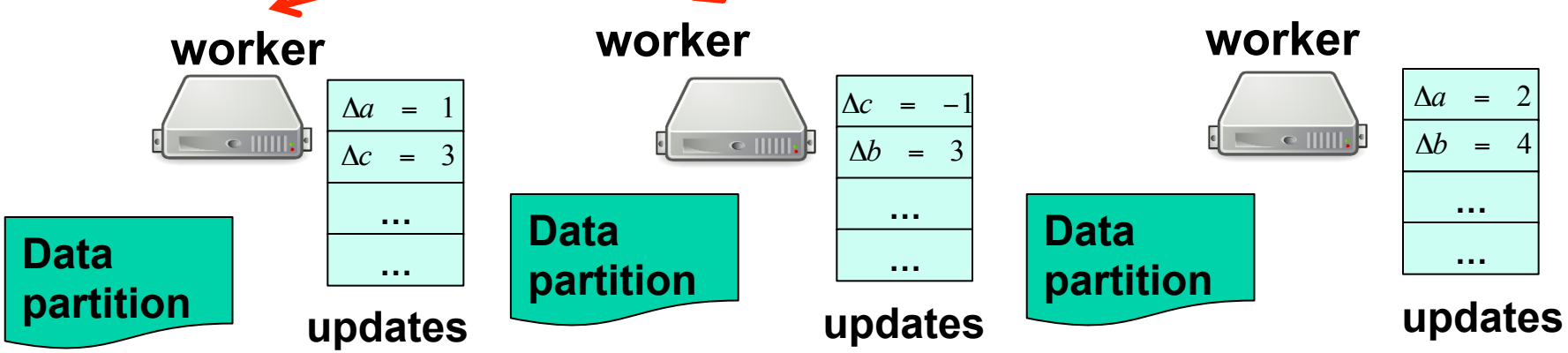
Goal: minimize total time to learn the model

3) Downstream tasks

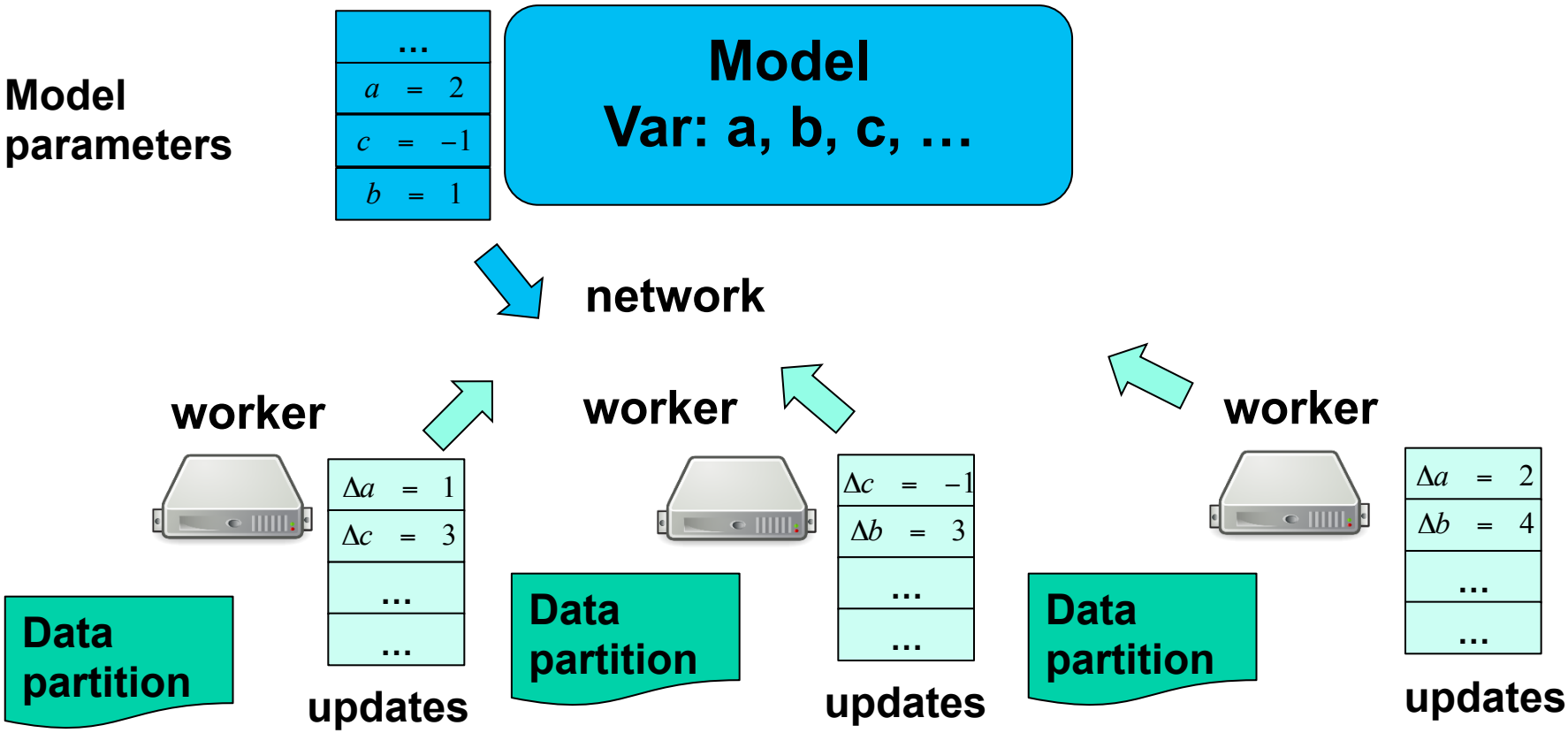
For example, "which movie should I, Netflix recommend to Bob?"

Data-Parallel, Iterative Training

$$Model^{(t+1)} = Model^{(t)} + \Delta(Model^{(t)}, Data)$$

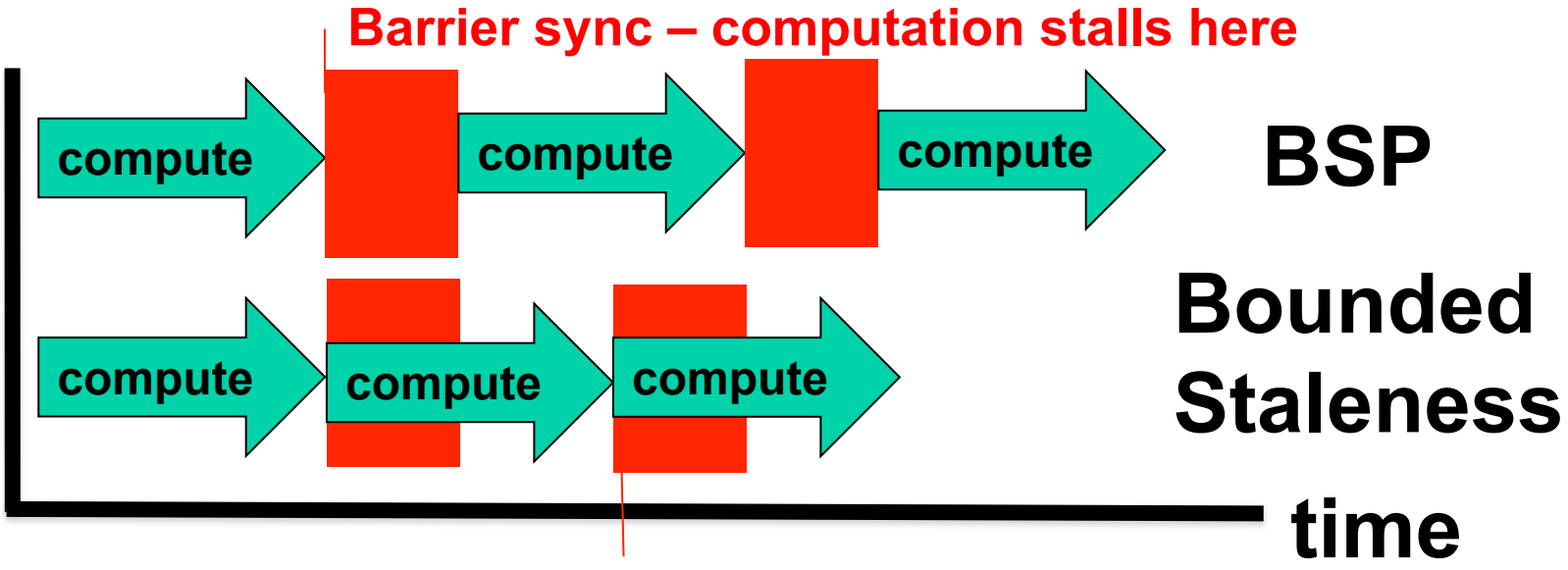


Data-Parallel, Iterative Training

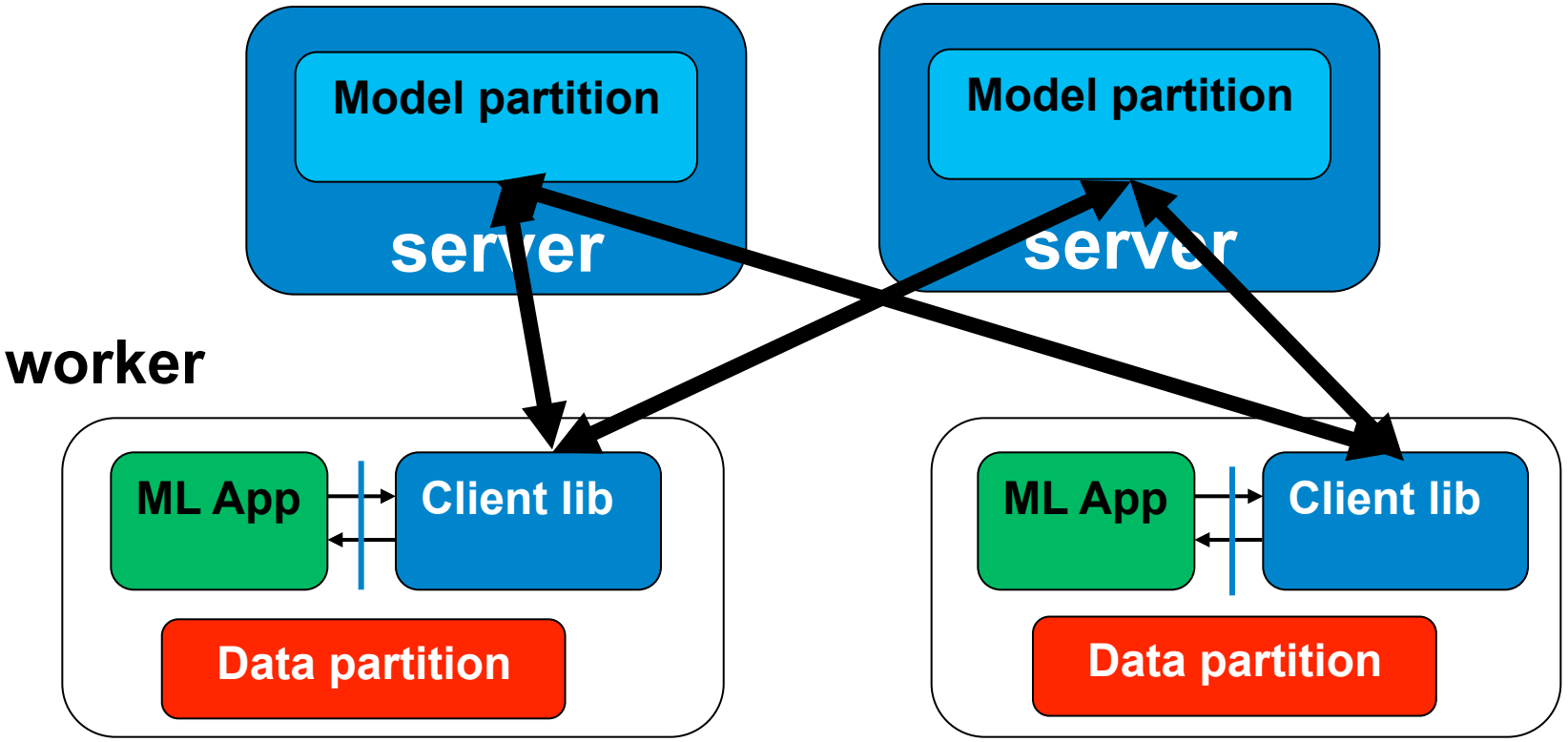


Faster Iterations From Relaxed Consistency

- **Bulk Synchronous Parallel (BSP) [Valiant'90]**
  - **Theoretically guaranteed convergence;**
  - **But may suffer from considerable synchronization overheads.**
- **Bounded Staleness [Ho'13] [Li'14]**
  - **ML theoretical convergence guarantees;**
  - **Also allows the synchronization overheads to be hidden.**
- **Computation uses local states, which might be stale**

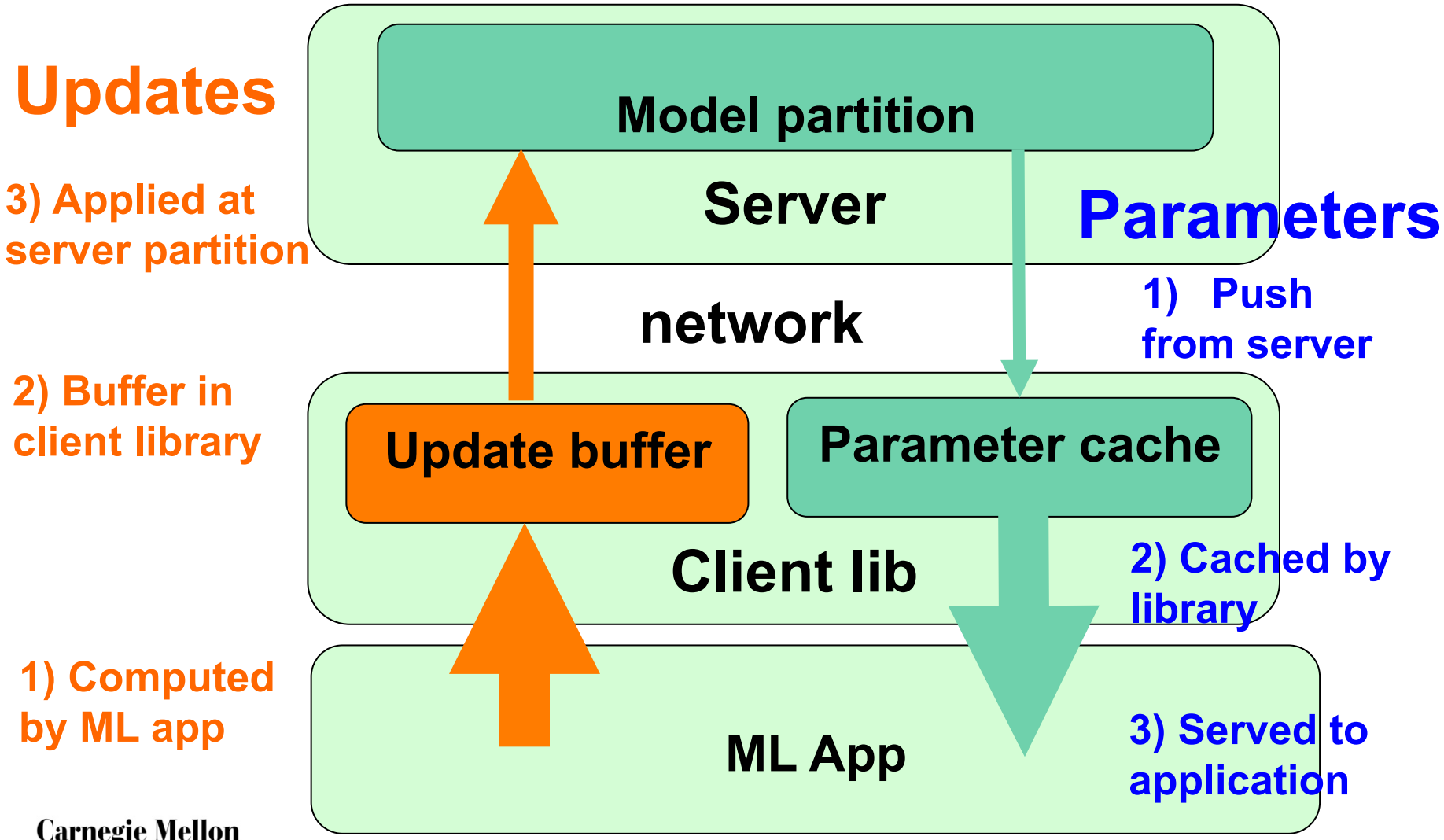


Bösen: A Key-value Store for ML



- Parameter Server [Power'10] [Ahmed'12] [Ho'13] [Li'14]
- Coherent shared memory abstraction for application
- Let the library worry about consistency, communication, etc

Communication in Bösen



Bounded Staleness + Continuous Comm

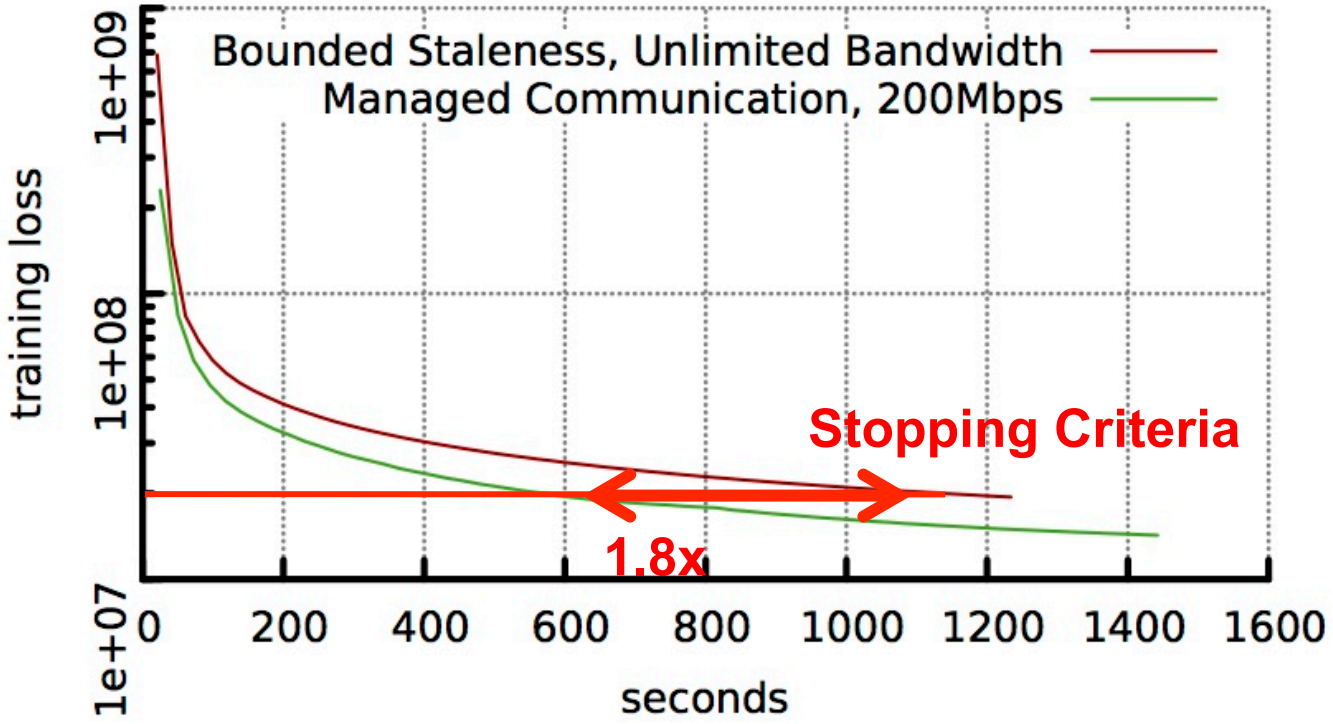
- Bounded Staleness only
  - Communicates only at iteration boundary
  - Ensures bounded staleness consistency

- Bounded Staleness + Managed Comm
  - Continuous communication
  - Same consistency guarantees as Bounded staleness
  - Fixed maximum per-node bandwidth

Messages are not equally important

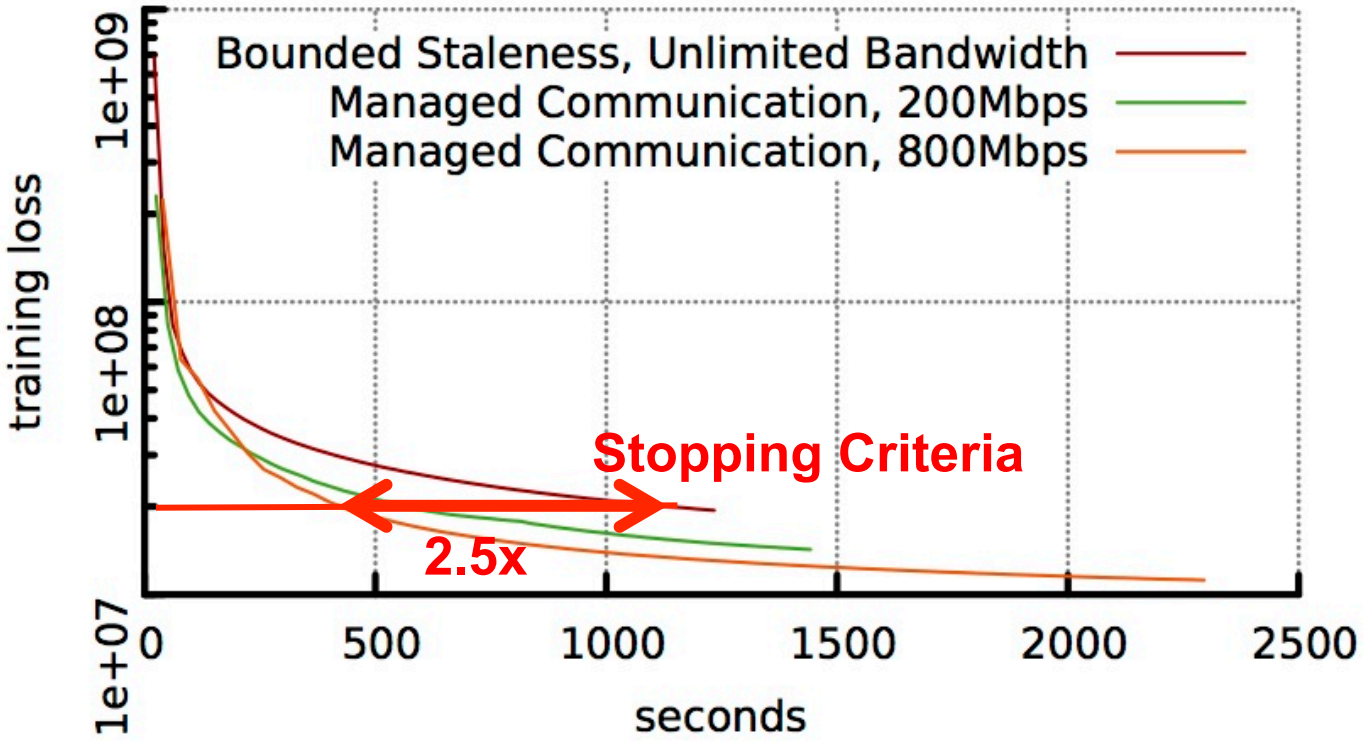
- Model parameters are not equally important
  - Majority of the parameters converge in a few iterations
- Communicate for more important ones
  - Magnitude of the changes indicates importance
- Exemplar prioritization policy:
  - Magnitude
  - Randomized

Quicker Communication Speeds Up Training



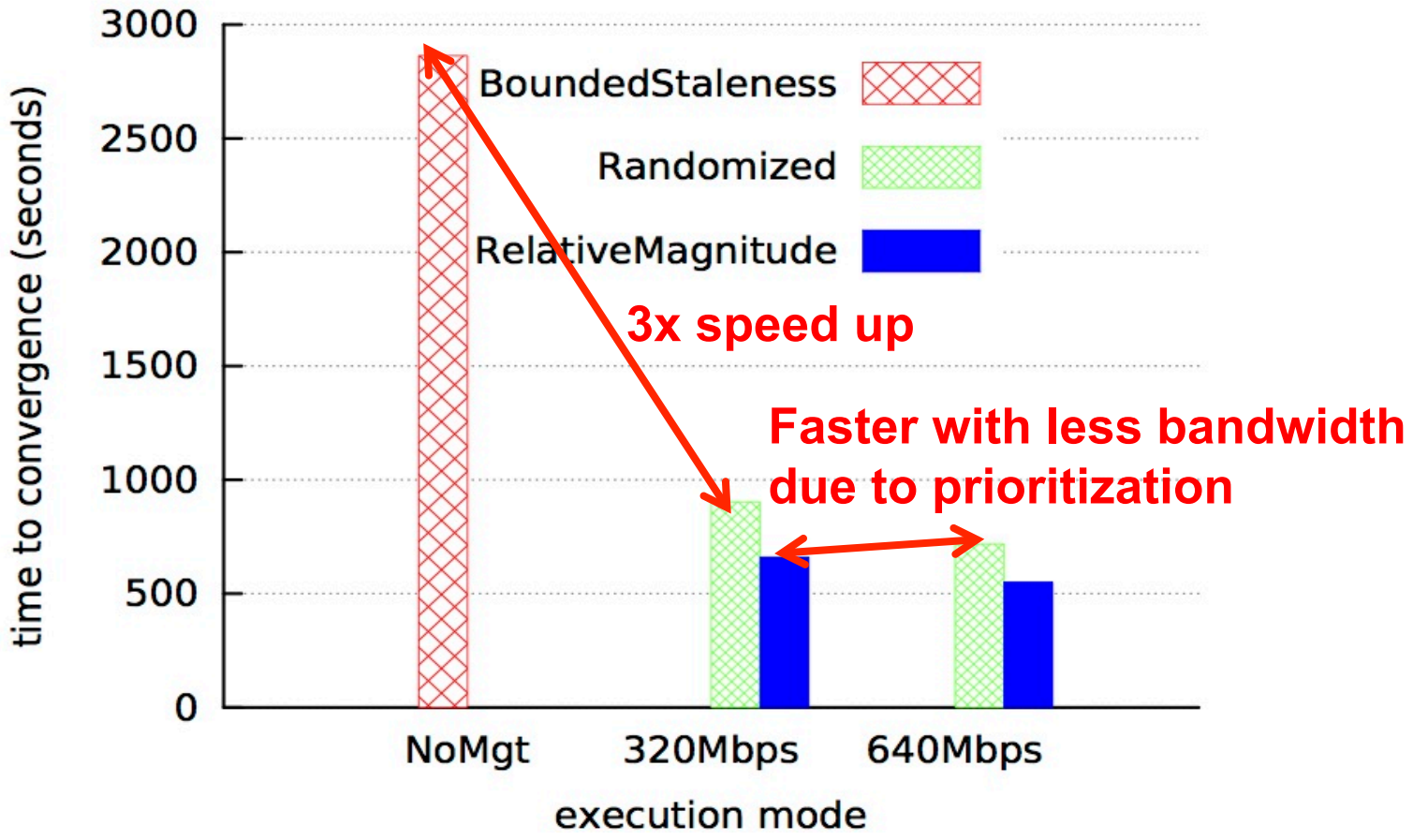
MF, 8x16 cores, 1GbE, Netflix data, rank=400.

Takes advantage of higher bandwidth



MF, 8x16 cores, 1GbE, Netflix data, rank=400.

Choosing What to Send Makes a Diff



LDA, NYTimes, # topics = 1000, 16x16 cores, 1GbE

Conclusion

- Bounded Staleness:

  Hide communication when you could, reduces progress per iteration

- Managed Communication:

  Communicate if there are advantages to be taken of

- Allocate network bandwidth according to importance of information

- 2-3X improvements for ML convergence rate

References

[Valiant'90] L. G. Valiant. *A bridging model for parallel computation*. *Commun. ACM*, 33(8):103–111, Aug. 1990.

[Ho'13] Q. Ho, J. Cipar, H. Cui, S. Lee, J. K. Kim, P. B. Gibbons, G. A. Gibson, G. Ganger, and E. P. Xing. *More effective distributed ml via a stale synchronous parallel parameter server*. NIPS, 2013.

[Li'14] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su. *Scaling distributed machine learning with the parameter server*. *OSDI*, 2014.

[Power'10] R. Power and J. Li. *Piccolo: Building fast, distributed programs with partitioned tables*. OSDI, 2010.

[Ahmed'12] A. Ahmed, M. Aly, J. Gonzalez, S. Narayanamurthy, and A. J. Smola. *Scalable inference in latent variable models*. WSDM, 2012*:*