# Scaling Up Distributed Data Analytics with Probabilistic Synchronous Machines

Liang Wang {liang.wang@cl.cam.ac.uk}

Computer Lab, University of Cambridge

## Problem

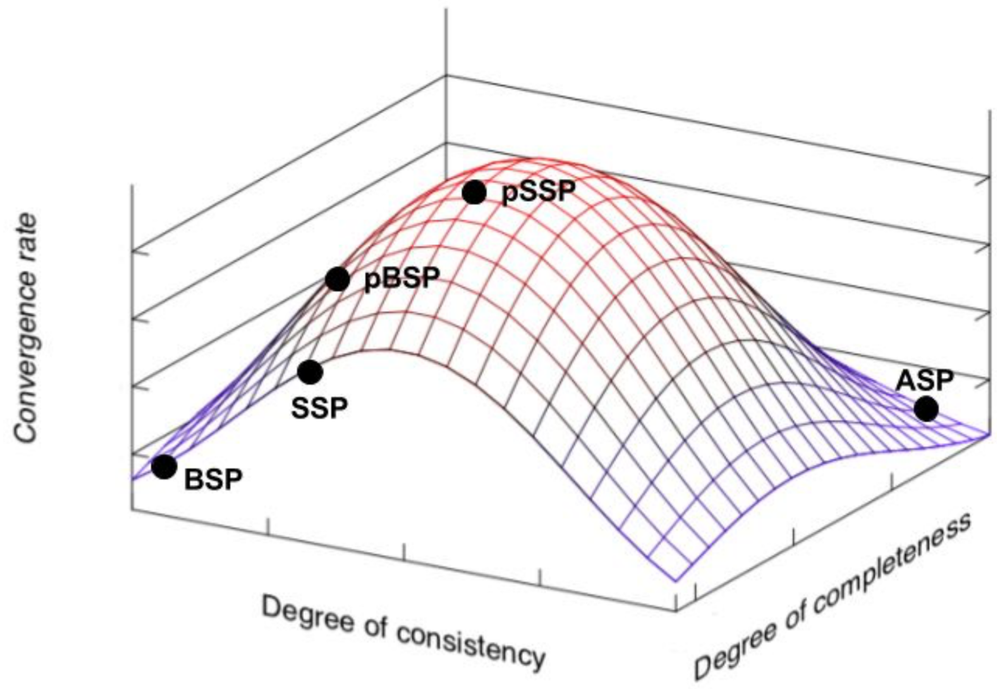
- Synchronisation plays a key role in many machine learning and neural network algorithms.
- Especially for those based on iterative algorithms such as Stochastic Gradient Descent, synchronisation improves the guarantees on convergence.
- There are several synchronous parallel machines widely used: BSP, SSP, ASP with different trade-off between iteration rate and consistency.
- This work aims to design a new synchronous parallel machine to scale up data analytics at edge networks.

## Proposed Solution: PSP

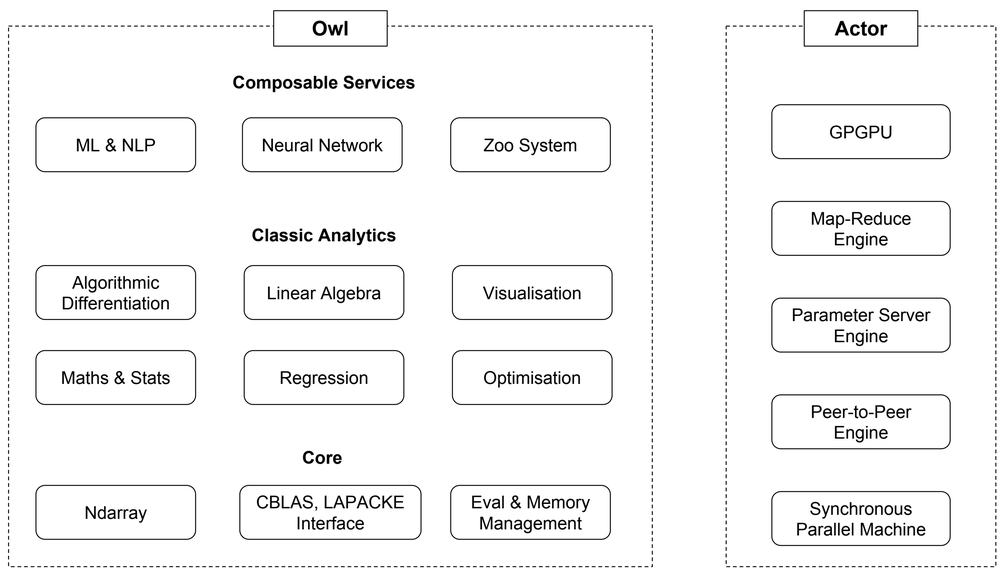
- Instead of maintaining the global state at a server, each node samples the network and applies previous synchronous parallel machines on the sample instead of the population.
- The nodes are organised into a structured overlay to guarantee the random sampling.
- Probabilistic Synchronous Parallel (PSP) is a higher order function over existing models.
- A new system primitive `sampling` is required but the new system is compatible with exsiting ones.

## Implications on System Design

- PSP adds a new dimension to the design space, i.e. Completeness, which allows us to find better synchronous parallel models.
- PSP allows us to decouple synchronisation with consistency, further derive fully distributed version of existing models.
- PSP transforms the server into a simple streaming server only dealing with the parameter updates, and it is easier to implement an efficient one.
- PSP is more robust against to stragglers and processes much better scalability than other models, therefore it is an ideal solution for edge networks.




## Owl + Actor Architecture




## Performance

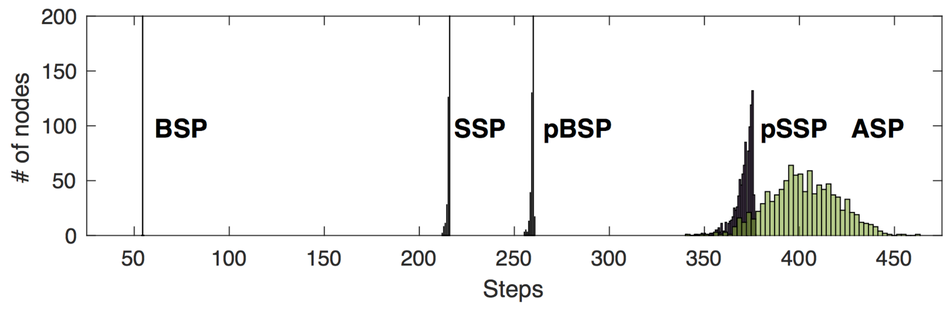



**Figure 1:** Progress distribution in steps

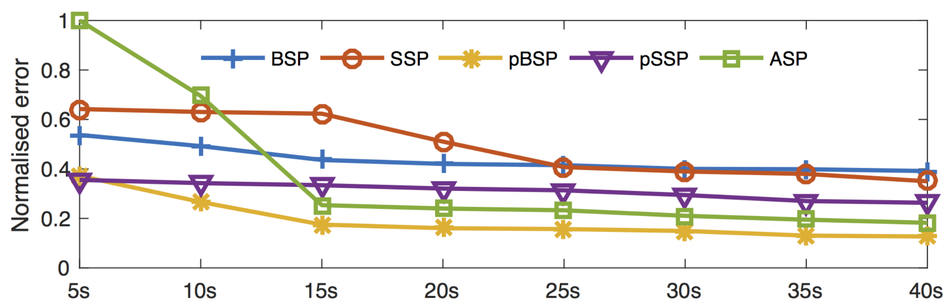



**Figure 2:** Normalised error values as a function of time

## Acknowledgements

This work is funded in part by the EPSRC NaaS (EP/K031724/2), Databox project (EP/N028260/1), and Contrive (EP/N028422/1).

## References

[1] Liang Wang. Owl: A general-purpose numerical library in ocaml. *CoRR*, abs/1707.09616, 2017.

[2] L. Wang, B. Catterall, and R. Mortier. Probabilistic Synchronous Parallel. *ArXiv e-prints*, 2017.

UNIVERSITY OF CAMBRIDGE

DATABOX