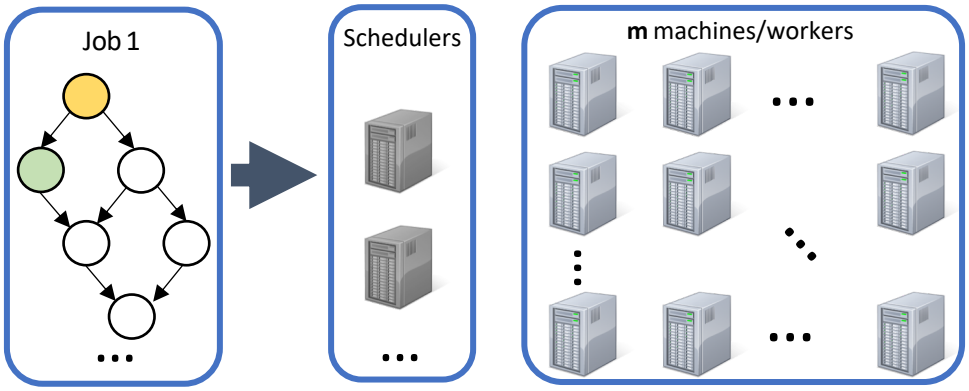
Using Scheduling to Accelerate AI

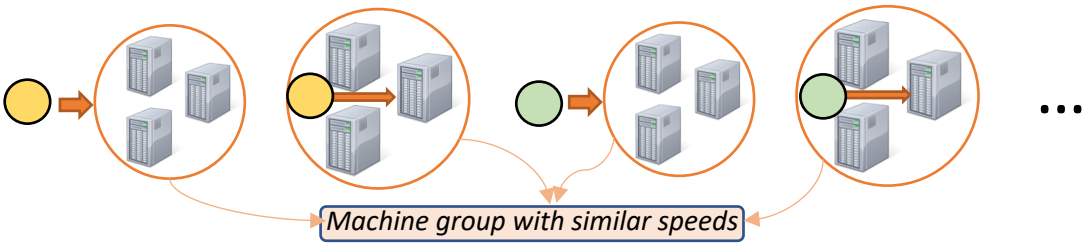
Yu Su[1] Xiaoqi Ren[1] Shai Vardi[2] Adam Wierman[1] Yuxiong He[3]
[1]Caltech [2]Purdue University [3]Microsoft

AI has proven to be a powerful tool in many areas, and finding ways to accelerate AI becomes very critical in making it accessible to the general audience. The key to the solution is to take advantage of massive computational resources by adopting efficient *parallelization* and the essence of such *parallelization* is *scheduling*.

Problem Setup



Algorithm



Step 1: group assignment based on machine speed

Step 2: assign the earliest task first

Repeat for tasks/nodes in Job 1

$$O(\log m / \log\log m)OPT^{(i)} + C \geq C_{max}(S)$$

Computational Time — *Communication Time* — *Schedule Length*

Take-away message

Our scheduling framework not only is able to handle complex graphs in large-scale production systems, but also comes with the first provable worst case theoretic bound. This can potentially facilitate efficient parallelization to democratize AI in the long run.