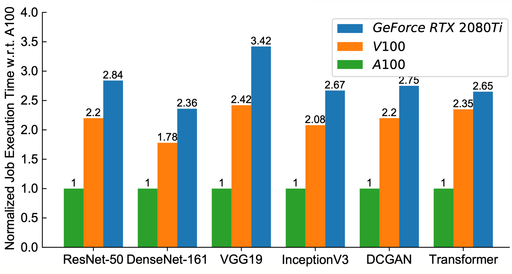
ISCAS 中国科学院软件研究所学术年会'2023 暨计算机科学国家重点实验室开放周
学术论文

异构GPU上面向截止时间与效率的深度学习任务调度
Hydra: Deadline-aware and Efficiency-oriented Scheduling for Deep Learning Jobs on Heterogeneous GPUs

(IEEE Transactions on Computers 2023) (CCF A)

杨紫超，吴恒，许源佳，吴悦文，钟华，张文博
联系人：杨紫超 邮箱：yangzichao21@otcaix.iscas.ac.cn

Deadline-aware and efficiency-oriented scheduling for deep learning jobs is challenging



Deep learning jobs have **diverse speeds** on heterogeneous GPUs.

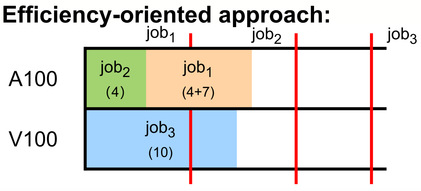
Name	V100	A100	Deadline
job_1	15	7	7
job_2	10	4	14
job_3	10	9	19

How to schedule jobs to improve global **efficiency** while meeting **deadlines**?

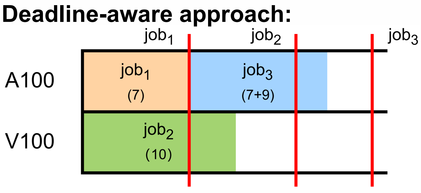
Limitations of existing scheduling approaches

Existing approaches typically prioritize either efficiency or meeting deadlines, without effectively leveraging quantitative analysis to integrate both factors, leading to **extended job completion time (JCT) or delays in meeting deadlines (tardiness)**:

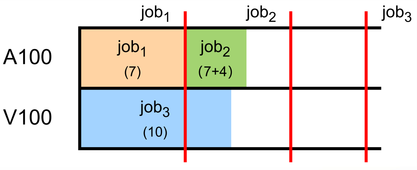
Efficiency-oriented approach:



Deadline-aware approach:



When consider both effiency and deadline missing, we can get the optimal:



How to design a new quantitative analysis approach to effectively support both objectives?

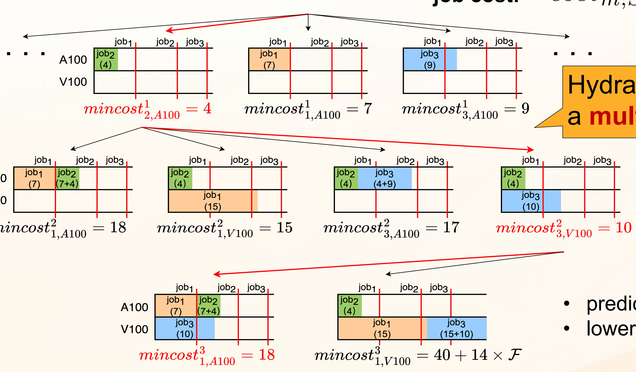
Hydra: A novel search methodology grounded in quantitative

Hydra effectively integrates a quantitative cost by amalgamating job completion time and tardiness:

job completion time: $c_n = w_n + e_n$

job tardiness: $tardiness_n = max\{0, c_n - d_n\}$

job cost: $cost_{m,S} = \sum_{j_n \in S} (c_n + \mathcal{F} \times tardiness_n)$



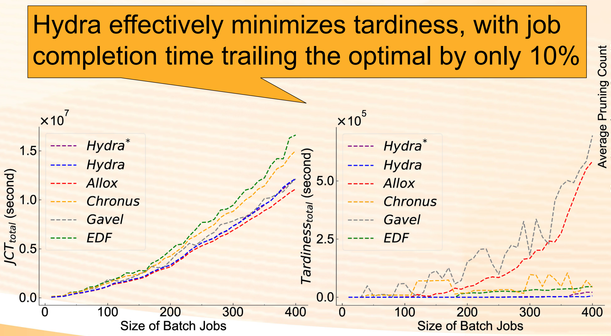
Hydra transforms the scheduling problem to a **multi-round min-cost search** problem.

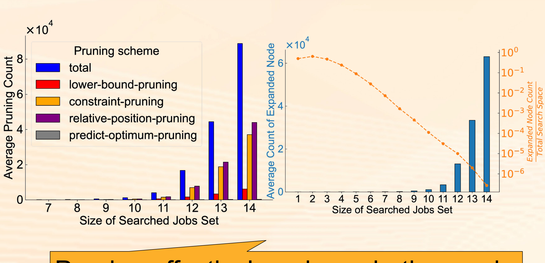
Min-cost search: an optimized search algorithm based on branch-and-bound, incorporating four pruning schemes:

- predict-optimum-pruning
- lower-bound-pruning
- relative-position-pruning
- constraint-pruning

Hydra guarantees deadline while improving more effiency

Hydra effectively minimizes tardiness, with job completion time trailing the optimal by only 10%





Pruning effectively reduces both search space and search overhead.