系统流程推演

场景： GPU 集群，假设目前有四台服务器，为了方便计算和描述，假设服务器资源是同构的，每台服务器有两个 w 类型的 worker 和一个 p 类型的 ps，$C_{sw}=2$，$C_{sp}=1$，$\forall s\in\{1,2,3,4\}$

输入任务及其相关配置：

三个任务$J=\{j_1,j_2,j_3\}$。对于每个任务，其配置如下：

（1） 任务 j 的处理能力，即在一个时隙内能够训练的 mini-batch 的数量：$k_j^{wp}=1$，如果任务 j 的所有 worker 和 ps 在同一台物理服务器上($\varphi_j=1$)；否则，$k_j^{wp}=0.5$，$\varphi_j=0$

（2） 三个任务的到达时间都是$r_j=3$，在第三个时隙分别到达，并假设在这之前系统中没有其他任务

（3） 训练数据集信息：训练轮次(epoch) $E_j=2$，数据块数量$D_{j_1}=D_{j_2}=2$，$D_{j_3}=3$，每个数据块的 mini-batch 数量$M_j=1$

（4） 任务效用函数$f_j(\hat{t}_j^l-r_j)=\frac{100\kappa}{1+e^{0.02(\hat{t}_j^l-r_j)}}$，$\kappa\in[1,5]$，此处令$\kappa=3$。$\hat{t}_j^l$是任务 j 使用调度 l 时的任务完成时间

（5） 任务 j 具有一个 ddl，$\tau_{j_1}=\tau_{j_2}=\tau_{j_3}=8$

（6） 由于场景是 GPU 集群，所以忽略了任务上传的时间

资源分配规则： 资源一旦分配，直到训练完再释放，不允许抢占

$$y_{jsw}(t)=y_{jsw}(t+1),\forall j,\forall s,\forall w,\forall t\in[a_j,\hat{t}_j-1]$$

$$z_{jsp}(t)=z_{jsp}(t+1),\forall j,\forall s,\forall p,\forall t\in[a_j,\hat{t}_j-1]$$

模拟调度过程：

（一）对于每个新到达的任务依次处理，为它们找到一个最优的调度方案并计算相应的任务效用值。

最开始时，系统中没有需要调度的任务，也不需要消耗资源，则系统中已经分配的 worker 资源$h_w(t)=\sum_s h_{sw}(t)=0$，PS 的使用量为$h_p(t)=\sum_s h_{sp}(t)=0$。此时，根据价格函数公式，worker 和 PS 的价格均为 0。

对于任务j_1和j_2，他们的开始时间范围是$a_j\in(3,8]$，分配给每个任务的 worker 的可能的数量为$D_w=\{1,2\}$（D_j数量限制）。对于每个参数组合$\{a_j,D_w\}$，分别尝试在分布式布局和集中式布局下计算最优调度方案。

例如，分布式策略中，对于任务j_1，当$\{a_{j_1}=4,D_w=1\}$时，任务的训练持续时间$d_{j_1}=\left\lceil\frac{E_jD_jM_j}{k_j^{wp}D_w}\right\rceil=8$，超过了调度区间(3,8]的长度，说明这种情况下无法产生可行方案。如果$\{a_{j_1}=4,D_w=2\}$，$d_{j_1}=\left\lceil\frac{E_jD_jM_j}{k_j^{wp}D_w}\right\rceil=4$，完成时间$\hat{t}_j^l=a_j+d_j=8$，任务$j_1$的最优调度的资源成本为 0，最优的任务效用$\mu_j=f_j(\hat{t}_j^l-r_j)-0\approx 142.85$。该元组参数在集中式策略下，$d_{j_1}=2$，$\hat{t}_j^l=6$，$\mu_j=f_j(\hat{t}_j^l-r_j)\approx 145.63$。所以系统实际上是按照集中式来部署的，也就是把服务器 s1（此时任意选）上的全部资源（两个 worker 和一个 ps）在$\forall t\in[4,6]$都给任务j_1使用

subject to:

$$\mu_j \geq f_j(\hat{\tau}_j^l - r_j) - \sum_{s\in[S]}\sum_{w\in[W]}\sum_{t\in[T]}\alpha_{sw}(t)\gamma_{jsw}^l(t) - \sum_{s\in[S]}\sum_{p\in[P]}\sum_{t\in[T]}\beta_{sp}(t)\zeta_{jsp}^l(t), \forall j, \forall l\in L_j \qquad (18a)$$

在安排完任务j_1后，此时，系统中的 worker 使用量更新$h_w(t)=\sum_s h_{sw}(t)=2$，PS 的使用量更新为$h_p(t)=\sum_s h_{sp}(t)=1$，所以根据以下价格函数设计，

$$\eta_w\big(h_w(t)\big) = (\theta_w)^{\frac{h_w(t)}{\sum_s \sum_w C_{sw}}} - 1$$

$$\theta_w = \max_j \left\{ \frac{f_j\big(\hat{t}_j^l - r_j\big)}{\sum_s \sum_w \sum_t y_{jsw}^l(t)} \right\} + 1$$

同理可得 PS 的价格函数

更新后的 worker 价格为$\eta_w(2) = (\frac{145.63}{2*2} + 1)^{\frac{2}{4*2}} - 1 \approx 2.5 - 1 = 1.5$，$\forall t \in [4,6]$

更新后的 PS 价格为$\eta_p(1) = (\frac{145.63}{1*2} + 1)^{\frac{1}{4*1}} - 1 \approx 3 - 1 = 2$，$\forall t \in [4,6]$

其他任务也以类似的方式计算。由于任务j_2的配置信息与j_1相似，所以可以获得的最优的调度方案是按照集中式的放置策略在$\forall t \in [6,8]$将服务器 s 上的全部资源（两个 worker 和一个 ps）分配给j_2使用，则资源成本为 0，$d_{j_2} = 2$，$\hat{t}_j^l = 8$，

$\mu_j = f_j\big(\hat{t}_j^l - r_j\big) - 0 \approx 142.85$

更新后的 worker 价格为$\eta_w(2) = (\frac{142.85}{2*2} + 1)^{\frac{2}{4*2}} - 1 \approx 2.5 - 1 = 1.5$，$\forall t \in [6,8]$

更新后的 PS 价格为$\eta_p(1) = (\frac{142.85}{1*2} + 1)^{\frac{1}{4*1}} - 1 \approx 3 - 1 = 2$，$\forall t \in [6,8]$

最后安排任务j_3，开始时间范围是$a_{j_3} \in (3,8]$，分配的 worker 的可能的数量为 $D_w = \{1,2,3\}$

分布式策略中，对于任务j_3，当$\{a_{j_3} = 4, D_w = 1 \text{ 或 } 2\}$时，任务的训练持续时间$d_{j_1} = \left\lceil \frac{E_j D_j M_j}{k_j^{wp} D_w} \right\rceil = 12 \text{ 或 } 6$，超过了调度区间(3,8]的长度，说明这种情况下无法产生可行方案。如果$\{a_{j_3} = 4, D_w = 3\}$，$d_{j_3} = \left\lceil \frac{E_j D_j M_j}{k_j^{wp} D_w} \right\rceil = 4$，完成时间$\hat{t}_j^l = a_j + d_j = 8$，任务$j_3$的最优调度的资源成本为$cost_{j_3} = 4 * 3 * 1.5 + 4 * 1 * 2 = 26, \forall t \in [4,8]$，最优的任务效用$\mu_j = f_j\big(\hat{t}_j^l - r_j\big) - cost_{j_3} \approx 116.85$。并且由于任务$j_3$必须使用 3 个 worker，而每台服务器只有两个 worker，所以一定是分布式的放置策略。比如，在$\forall t \in [4,8]$把服务器 s3 上的全部资源（两个 worker 和一个 ps）以及 s4 的一个 worker 都给任务j_3使用。

（二）分布式/集中式放置策略下任务资源成本最小的分配方案--COST_C 和 COST_D 算法

首先枚举参数组合$\{a_j, D_w\}$，计算任务的训练持续时间d_j，并计算任务完成时间$\hat{t}_j^l$。如果完成时间超过了任务的 ddl，或者剩余的可用 worker 不能满足任务需求，则停止分配。

对于分布式的部署：选择可以使用最小数量的 ps 来满足带宽约束的服务器来放置所需的 ps。此外，ps 的分配不能违反容量约。对于剩余的 worker，从资源最少的服务器开始放置它们。通过这种方式，保留了足够资源的服务器，从而增加了集中放置的可能性

比如：对于任务j_1，当$\{a_{j_1} = 4, D_w = 2\}$，$d_{j_1} = \left\lceil \frac{E_j D_j M_j}{k_j^{wp} D_w} \right\rceil = 4$，完成时间$\hat{t}_j^l = a_j + d_j = 8$，并且算法限制在任务训练过程中不允许抢占。分布式放置中，至少有一个 worker 和 PS 没有部署在同一台物理服务器上，需要找到能够部署 PS 的目标服务器。

首先计算每台服务上可用 worker 的最大数量：

$$D_{sw} = \min_{t \in [a_j, t_j]} \{D_w - 1, C_{sw} - h_{sw}(t)\} = \min_{t \in [4,8]} \{2 - 1, 2 - 0\} = 1$$

每台服务上可用 ps 的最大数量：

$$D_{sp} = \min_{t \in [a_j, t_j]} \left\{ \left\lceil \frac{(D_w - D_{sw}) b_w}{b_p} \right\rceil, C_{sp} - h_{sp}(t) \right\} = \min_{t \in [4,8]} \left\{ \left\lceil \frac{(2-1) * 1}{1} \right\rceil, 1 - 0 \right\} = 1$$

现有的四台服务器都是合格的，即$H = \{h_1, h_2, h_3, h_4\}$。先把服务器 s1 中的一个 worker 和一个 PS 分配给任务j_1。剩下的一个 worker 是从其他服务器 s2, s3, s4 中选择，根据可用 worker 的数量按非递减顺序排序。

此时资源价格为 0，所以相应的成本为 0。那么理想的收入为$f_j\left(\hat{t}_j^l - r_j\right) \approx 142.85$，并且任务效用$\mu_{j_1} = 142.85$

安排完任务j_1，消耗的 worker 数量为$h_w(t)=2$， 消耗的 PS 量为$h_p(t)=1$

更新后的 worker 价格为$\eta_w(2)=(\frac{142.85}{4*2}+1)^{\frac{2}{4*2}}-1\approx 2-1=1$，$\forall t\in[4,8]$

更新后的 PS 价格为$\eta_p(1)=(\frac{142.85}{4*1}+1)^{\frac{1}{4*1}}-1\approx 2.5-1=1.5$，$\forall t\in[4,8]$

注意，以上调度只是模拟调度，实际最优调度由最优调度算法 Alg.bestsche 确定。假设上述调度实际上被算法 Alg.bestsche 所采用，那么通过重复上述过程，可以为任务j_2分配服务器 s3 的一个 worker 和一个 PS，服务器 s4 的一个 worker.

资源成本：$cost_{j_2}=4*2*1+4*1*1.5=14$

理想的收入：$f_j\left(\hat{t}_j^l-r_j\right)\approx 142.85$

任务效用：$\mu_{j_2}=f_j\left(\hat{t}_j^l-r_j\right)-cost_{j_2}=128.85>0$

因此这也是一种可行的调度方案，但不是最优的

如果还要安排任务j_3，继续更新资源使用量：消耗的 worker 数量为$h_w(t)=4$，消耗的 PS 量为$h_p(t)=2$

但是θ_w是不变的，对于j_1和j_2，$f_j\left(\hat{t}_j^l-r_j\right)$和$\sum_s\sum_w\sum_t y_{jsw}^l(t)$都是一样的

$$\theta_w=\max_{j\in[J]}\left\{\frac{f_j\left(\hat{t}_j^l-r_j\right)}{\sum_s\sum_w\sum_t y_{jsw}^l(t)}\right\}+1$$

但是价格更新为：

更新后的 worker 价格为$\eta_w(4)=(\frac{142.85}{4*2}+1)^{\frac{4}{4*2}}-1\approx 4.3-1=3.3$，$\forall t\in[4,8]$

更新后的 PS 价格为$\eta_p(2)=(\frac{142.85}{4*1}+1)^{\frac{2}{4*1}}-1\approx 6-1=5$，$\forall t\in[4,8]$

此时按照每台服务器已经分配的资源量的非递减顺序排序（空闲资源量最多的排在最前面）

那么此时服务器的排序为$\{s4,s2,s3,s1\}$，当$\left\{a_{j_3}=4,D_w=3\right\}$时，$d_{j_3}=\left\lceil\frac{E_jD_jM_j}{k_j^{wp}D_w}\right\rceil=4$，完成时间$\hat{t}_j^l=a_j+d_j=8$

计算每台服务上可用 worker 的最大数量：

$$D_{sw}=\min_{t\in[a_j,t_j]}\{D_w-1,C_{sw}-h_{sw}(t)\}=\min_{t\in[4,8]}\{3-1,2-1\}=1$$

每台服务上可用 ps 的最大数量：

$$D_{sp}=\min_{t\in[a_j,t_j]}\left\{\left\lceil\frac{(D_w-D_{sw})b_w}{b_p}\right\rceil,C_{sp}-h_{sp}(t)\right\}=\min_{t\in[4,8]}\left\{\left\lceil\frac{(3-1)*1}{1}\right\rceil,1-0\right\}=1$$

先把服务器 s4 中的一个 worker 和一个 PS 分配给任务j_3。剩下的两个 worker 是从其他服务器 s2, s3 中选择，根据可用 worker 的数量按非递减顺序排序。

资源成本：$cost_{j_3}=4*3*3.3+4*1*5=59.6$

理想的收入：$f_j\left(\hat{t}_j^l-r_j\right)\approx 142.85$

任务效用：$\mu_{j_2}=f_j\left(\hat{t}_j^l-r_j\right)-cost_{j_2}=83.25>0$

此后再去更新资源使用量：消耗的 worker 数量为$h_w(t)=7$， 消耗的 PS 量为$h_p(t)=3$

θ_w取最大值，所以$f_j\left(\hat{t}_j^l-r_j\right)$越大，$\sum_s\sum_w\sum_t y_{jsw}^l(t)$越小，比值越大，所以此时$\theta_w$还是不变的

更新后的 worker 价格为$\eta_w(7)=(\frac{142.85}{4*2}+1)^{\frac{7}{4*2}}-1\approx 13-1=12$，$\forall t\in[4,8]$

更新后的 PS 价格为$\eta_p(3)=(\frac{142.85}{4*1}+1)^{\frac{3}{4*1}}-1\approx 15-1=14$，$\forall t\in[4,8]$

（三）实际的最优调度

（1）任务j_1的最优调度方案：集中式放置且$\left\{a_{j_1}=4,D_w=2\right\}$

$d_{j_1}=\left\lceil\frac{E_jD_jM_j}{k_j^{wp}D_w}\right\rceil=2$，完成时间$\hat{t}_j^l=a_j+d_j=6$

资源成本：$cost_{j_1}=0$，价格为 0，$\forall t\in[4,6]$

理想的收入：$f_j\left(\hat{t}_j^l-r_j\right)\approx 145.63$

任务效用：$\mu_j=f_j\left(\hat{t}_j^l-r_j\right)-0=145.63$

worker 使用量更新$h_w(t)=\sum_s h_{sw}(t)=2$

PS 的使用量更新为$h_p(t)=\sum_s h_{sp}(t)=1$

更新后的 worker 价格为$\eta_w(2)=(\frac{145.63}{2*2}+1)^{\frac{2}{4*2}}-1\approx 2.5-1=1.5$，$\forall t\in[4,6]$

更新后的 PS 价格为$\eta_p(1) = (\frac{145.63}{1*2} + 1)^{\frac{1}{4*1}} - 1 \approx 3 - 1 = 2$，$\forall t \in [4,6]$

（2）任务j_2的最优调度方案：集中式放置且$\{a_{j_2} = 6, D_w = 2\}$

$d_{j_2} = \left\lceil \frac{E_j D_j M_j}{k_j^{wp} D_w} \right\rceil = 2$，完成时间$\hat{t}_j^l = a_j + d_j = 6 + 2 = 8$

资源成本：$cost_{j_2} = 0$，价格为 0，$\forall t \in [6,8]$

理想的收入：$f_j(\hat{t}_j^l - r_j) \approx 142.85$

任务效用：$\mu_j = f_j(\hat{t}_j^l - r_j) - 0 = 142.85$

worker 使用量更新$h_w(t) = \sum_s h_{sw}(t) = 2$

PS 的使用量更新为$h_p(t) = \sum_s h_{sp}(t) = 1$

更新后的 worker 价格为$\eta_w(2) = (\frac{142.85}{2*2} + 1)^{\frac{2}{4*2}} - 1 \approx 2.5 - 1 = 1.5$，$\forall t \in [6,8]$

更新后的 PS 价格为$\eta_p(1) = (\frac{142.85}{1*2} + 1)^{\frac{1}{4*1}} - 1 \approx 3 - 1 = 2$，$\forall t \in [6,8]$

（3）任务j_3的最优调度方案：分布式放置且$\{a_{j_3} = 4, D_w = 3\}$

$d_{j_3} = \left\lceil \frac{E_j D_j M_j}{k_j^{wp} D_w} \right\rceil = 4$，完成时间$\hat{t}_j^l = a_j + d_j = 4 + 4 = 8$

资源成本：$cost_{j_3} = 4 * 3 * 1.5 + 4 * 1 * 2 = 26$，$\forall t \in [4,8]$

理想的收入：$f_j(\hat{t}_j^l - r_j) \approx 142.85$

任务效用：$\mu_j = f_j(\hat{t}_j^l - r_j) - 26 = 116.85$

worker 使用量更新$h_w(t) = \sum_s h_{sw}(t) = 5, \forall t \in [4,8]$

PS 的使用量更新为$h_p(t) = \sum_s h_{sp}(t) = 2, \forall t \in [4,8]$

更新后的 worker 价格为$\eta_w(5) = (\frac{145.63}{2*2} + 1)^{\frac{5}{4*2}} - 1 \approx 9.6 - 1 = 8.6$，$\forall t \in [4,8]$

更新后的 PS 价格为$\eta_p(2) = (\frac{145.63}{1*2} + 1)^{\frac{2}{4*1}} - 1 \approx 8.6 - 1 = 7.6$，$\forall t \in [4,8]$

根据最优调度方案，计算出的目标最优值为：

$\sum_{j \in [J]} f_j(\hat{t}_j - r_j)$=$145.63 + 142.85 + 142.85 = 431.33$