1 Введение

В ходе исполнения задачи в распределенной вычислительно среде часто встает задача пересылки одинакового объема данных нескольким узлам кластера. В зависимости от конфигурации, связности и прочих характеристик коммуникацонной среды необходимо применять те или иные модели передачи данных. Требуют отдельного рассмотрения случаи тесносвязного кластера и метакластера с неоднородной коммуникационной средой. Далее приведен анализ различных моделей для этих случаев.

2 Оптимизация широковещательных рассылок в однородных средах.

2.1 Введение

Ниже будут рассмотрены некоторые методы распределения данных по однородному вычислительному кластеру и проанализирована их производительность.

2.2 Постановка и моделирование

Требуется передать файл с исходного узла на n принимающих, размер файла может быть весьма велик и доходить до десятков гигабайт. Передача может вестись не только с исходной узла (*source*), возможна ретрансляция узлами-приёмниками. Таким образом, строится некоторый ориентированный граф, охватывающий все узлы-приёмники. Ниже будет представлено несколько различных типов передающих сетей. Общие для всех моделей обозначения:

N_{dest} — количество принимающих узлов.

B_{disk} — скорость чтения единицы данных из локальной дисковой памяти.

L_{disk} — задержка перед чтением любого объёма данных из локальной дисковой памяти.

B_{net} — скорость чтения единицы данных по сетевому соединению.

L_{net} — задержка перед чтением любого объёма данных по сетевому соединению.

B_{total} — скорость передачи данных в построенной коммуникационной сети. Можно взять равной средней скорости выхода данных с исходного узла по каждому соединению.

Для простейшей ситуации, когда $N_{dest} = 1$, скорость передачи можно считать равной $B_{total} = \min(B_{disk}, B_{net})$.

2.3 Передача от одного всем

Простейшая модель, когда исходный узел передаёт каждому принимающему непосредственно, без ретрансляций (рис. 1). Эта схема отличается простотой, однако для неоднородных коммуникационных сетей она не применима. Её отличает простота реализации, однако в случае, если $B_{net} < B_{disk} \cdot N_{dest}$, начнутся значительные потери производительности. Скорость B_{total} в такой ситуации можно считать равной $B_{total} = \min(B_{net}/N_{dest}, B_{disk})$.

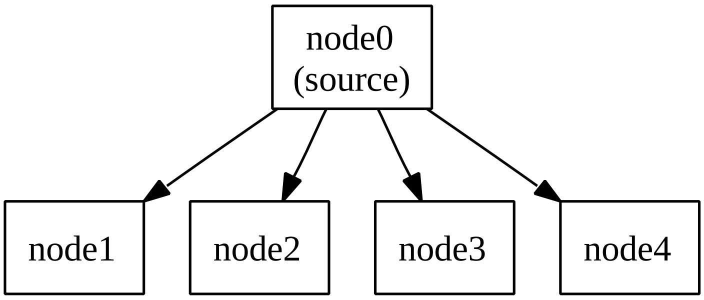

Рис. 1: Передача от одного ко всем. $N_{dest} = 4$

2.4 Последовательная передача

В этой модели блоки данных передаются по цепочке узлов (рис. 2). Каждый ретранслирующий узел записывает полученный блок на локальный диск и передаёт следующему в цепочке. Эта схема значительно разгружает сетевую инфраструктуру, так как ни один узел не имеет более одного входящего и одного исходящего потока. Таким образом, на сетях с коммутируемыми каналами данные передачи будут идти практически независимо. Кроме этого, в случае гетерогенной коммуникационной сети возможно так расположить узлы в последовательность, что по наиболее узким местам произойдёт только одна передача. Время, затраченное на передачу блока равно времени затраченному на запись блока на диск и пересылку по сети. Таким образом, для линейной передачи пропускная способность $B_{total} = \frac{B_{net} \cdot B_{disk}}{B_{net}+B_{disk}}$.

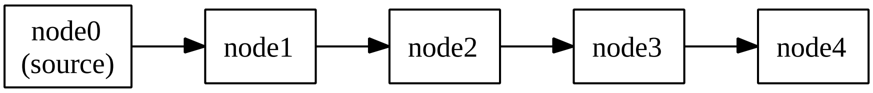

Рис. 2: Последовательная передача. $N_{dest} = 4$

2.5 Передача по двоичному дереву

Эта модель построена аналогично последовательной передаче, однако некоторые узлы дополнительно передают блоки данных двум другим (рис. 3). Занумеруем узлы каким-либо образом, установив исходному узлу индекс 0. Тогда по данным индексам построим двоичное дерево с корнем в узле с нулевым индексом. Построение происходит по двоичной записи индекса — у $node_k$ источником является $node_{k_R}$, а передача идёт на $node_{k_L}$ и $node_{k_L|1}$, где k_R — k побитовый сдвиг вправо, а k_L — влево. Таким образом, для любого N_{dest} мы построим дерево глубины не более $[\log N_{dest}] + 1$. Вычисление скорости передачи аналогично последовательному случаю $B_{total} = \frac{B_{net} \cdot B_{disk}}{B_{net}+B_{disk}}$. Также, данная схема может быть адаптирована под гетерогенные коммуникационные сети, если при нумерации узлов учитывать информацию о топологии вычислительной системы.

2.6 Сравнительные результаты

В ходе тестов передача данных шла с использованием коммуникационной среды SCI с пропускной способностью около 240Mb/s и жёсткие диски со скоростью около 40Mb/s. Передача данных

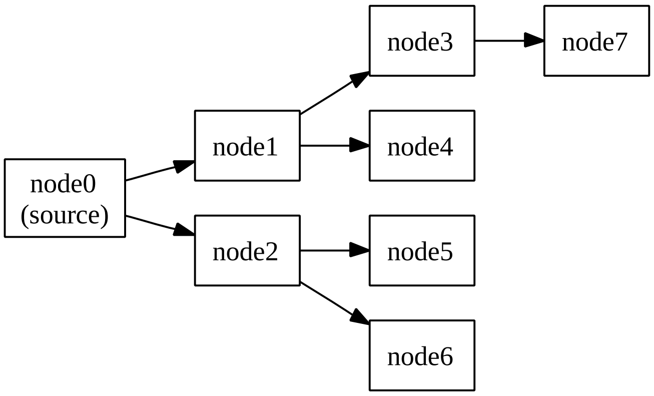

Рис. 3: Передача по двоичному дереву. $N_{dest} = 7$

по сетевой среде производилась посредством средств MPI, что объясняет некоторые странности в полученных результатах. Для модели передачи по дереву были достигнуты скорости $B_{total} = 34.5\text{Mb/s}$ при $N_{dest} = 7$. При аналогичных условиях для последовательной пересылки и для передачи от одного всем были получены скорости $B_{total} = 28.5\text{Mb/s}$ и $B_{total} = 19.5\text{Mb/s}$ соответственно. На графике (рис. 4) представлены усреднённые результаты тестов на файлах размером от 2 до 16 Gb по трём рассмотренным выше моделям коммуникационных сетей.

3 Оптимизация широковещательных рассылок в неоднородных средах.

3.1 Постановка задачи.

Рассмотренные выше модели распределения данных подходят только для однородных коммуникационных сред, так как не учитывают сложность передачи между компонентами связности метакластера. Ниже будет предложен простой алгоритм построения графа пересылок и доказательство его оптимальности.

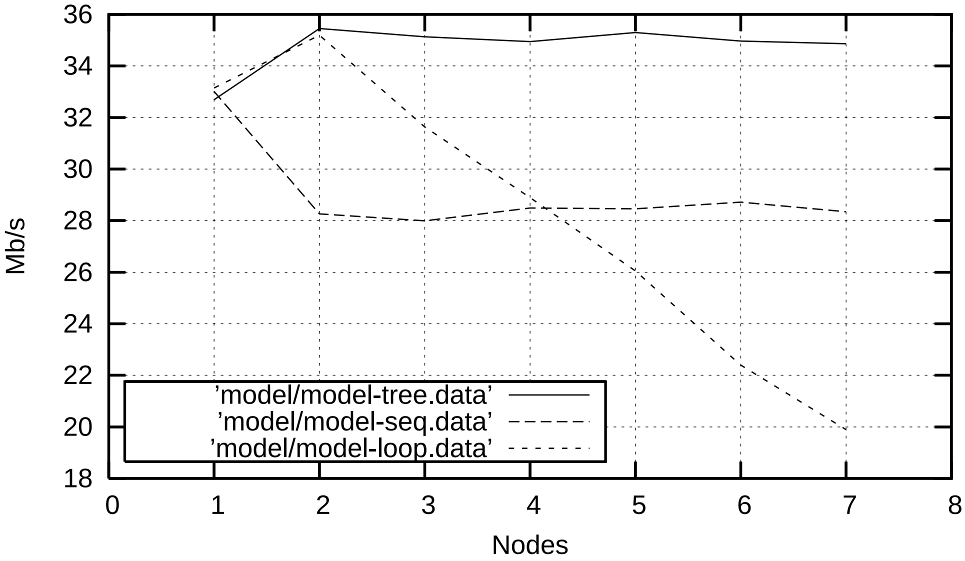

Рис. 4: Скорость передачи B_{total} в зависимости от количества узлов. $N_{dest} = 1 - 7$

3.2 Формализация задачи.

Дан полный граф Γ с вершинами $n(\Gamma) = \{n_0, \dots, n_N\}$, вершина n_0 и весовая функция на ребрах $f : n(\Gamma) \times n(\Gamma) \to \mathbb{R}$. Для ребра $e = (n_i, n_j)$ положим $f(e) = f(n_i, n_j)$. Далее везде f предполагается симметричной, то есть $f(x, y) = f(y, x)\ \forall x, y \in n(\Gamma)$.

Определение 1. Весом графа Γ' с ребрами $e(\Gamma') = \{e_1, \dots, e_N\}$ называется функция равная $\mu(\Gamma') = \sum_{1 \dots N} f(e_i)$.

Определение 2. Граф Γ' называется связным, если для любого $n_i \in n(\Gamma)$ существует путь по ребрам $e^i_1, \dots, e^i_k$ из n_0 в n_i

Определение 3. Связный граф Γ' называется оптимальным, если его вес минимален среди всех связных графов Γ'' с теми же вершинами.

Требуется построить оптимальный граф Γ' с вершинами $n(\Gamma') = n(\Gamma) = \{n_0, \dots, n_N\}$. Заметим, что оптимальных графов может быть много, например в случае если $\forall i, j\ f(n_i, n_j) = \text{const}$, оптимальным будет любой связный граф.

3.3 Построение.

Будем строить Γ' по индукции. База: $\Gamma'_0 = (n'_0, \varnothing)$ — единственный граф из одной вершины и поэтому оптимальный.

Шаг: пусть имеется оптимальный граф Γ'_k , построим Γ'_{k+1} с теми же свойствами. Пусть $n(\Gamma'_k) = \{n'_0 = n_0, n'_1, \ldots, n'_k\}$ — множество вершин Γ'_k, найдем вершины $n'_{k+1} \in n(\Gamma) \setminus n(\Gamma'_k)$ и $n'_j \in n(\Gamma'_k)$ такие, что $\forall \hat{n}' \in n(\Gamma'_k), \forall \hat{n} \in n(\Gamma) \setminus n(\Gamma'_k)$ $f(n'_j, n'_{k+1}) \leqslant f(\hat{n}', \hat{n})$. То есть, n'_{k+1} — вершина из $\Gamma \setminus \Gamma'_k$, такая что "расстояние" от нее до Γ'_k минимально среди всех $n(\Gamma) \setminus n(\Gamma'_k)$

Будем нумеровать ребра в порядке добавления в Γ'_{k+1}, то есть, $\Gamma'_{m+1} = \Gamma'_m \cup \{e_m\} \cup \{n'_m\}$.

Утверждение. Γ'_{k+1} *оптимален.*

Рассмотрим произвольный связный граф без циклов $\Gamma'' = \Gamma''_{k+1}$ на вершинах $n(\Gamma'_{k+1})$. Покажем что $\mu(\Gamma'_{k+1}) \leqslant \mu(\Gamma'')$.

Определение. Положим характеристику $\chi_{n'_l}(\hat{\Gamma})$ равной количеству ребер в графе $\hat{\Gamma}$ имеющих одним из концов вершину n'_l.

Пусть $n'_{l_1}, n'_{l_2}, \ldots, n'_{l_q},\ q = \chi_{n'_{k+1}}(\Gamma'')$ — вершины, соединенные с n'_{k+1} ребром в графе Γ'', выписаные в порядке добавления их в граф Γ'_{k+1}, то есть, $\forall j \in \{1, \ldots, q-1\}\ \exists i\ n'_{l_j} \in n(\Gamma'_i),\ n'_{l_{j+1}} \notin n(\Gamma'_i)$. Далее рассмотрим два случая.

$\chi_{n'_{k+1}}(\Gamma'') = 1$. Тогда $f(n'_{l_1}, n'_{k+1}) \geqslant f(n'_j, n'_{k+1})$, где n'_j — вершина, соединенная с n'_{k+1} ребром в графе Γ'_{k+1}, а граф Γ'_k оптимален по предположению индукции.

$\chi_{n'_{k+1}}(\Gamma'') > 1$. Построим граф $\hat{\Gamma}''$ с весом $\mu(\hat{\Gamma}'') \leqslant \mu(\Gamma'')$ такой, что $\chi_{n'_{k+1}}(\hat{\Gamma}'') = \chi_{n'_{k+1}}(\Gamma'') - 1$ Положим $i_1 = l_q$ и $i_2 = l_{q-1}$. Рассмотрим ребро $\hat{e}_2 = (n'_{k+1}, n'_{i_2})$. Так как Γ'' не имеет циклов, после удаления $\hat{e}_2$ он распадется на две компоненты связности. Обозначим Γ''_+ компоненту содержащую n'_{i_2} и Γ''_- — оставшуюся часть.

Утверждение 1. Для любой вершины $n'_l \in n(\Gamma'_{k+1})$, ребра $e_m \in e(\Gamma'_{k+1})$, выходящего из нее, и любого ребра $e_i \in e(\Gamma'_{k+1}),\ l < i \leqslant m$ верно, что $f(e_i) \leqslant f(e_m)$.

Утверждение 2. Для любой вершины $n'_l \in n(\Gamma'_{k+1})$, ребра $e \notin e(\Gamma'_{k+1})$ или $e = e_{k+1}$, выходящего из нее, и любого ребра $e_i \in e(\Gamma'_{k+1}),\ l < i \leqslant k+1$ верно, что $f(e_i) \leqslant f(e)$.

Определение. Назовем ребро $e = (n_i, n_k)$ *соединяющим*, если $e \in e(\Gamma'_k)$ и $n_i \in n(\Gamma''_+)$, $n_j \in n(\Gamma''_-)$ или $n_j \in n(\Gamma''_+)$, $n_i \in n(\Gamma''_-)$.

Предположим что найдется соединяющее ребро $e_{j'_2}$, такое что $i_2 < j'_2 \leqslant i_1$, тогда, $f(e_{j'_2}) \leqslant f(e)$ по утверждению 2. Рассмотрим граф $\hat{\Gamma}'' = e_{j'_2} \cup \Gamma'' \setminus e$, для него верно, что $\mu(\hat{\Gamma}'') \leqslant \mu(\Gamma'')$ и $\chi_{n'_{k+1}}(\hat{\Gamma}'') = \chi_{n'_{k+1}}(\Gamma'') - 1$, таким образом построен граф с искомыми свойствами. Иначе, рассмотрим вершину $n'_{i_3} \in \Gamma'_-$ такую, что из нее выходит ребро e_{j_3}, $j_3 > i_2$. Тогда, $f(e_{j_3}) \leqslant f(e)$ по утверждению 1. Если найдется соединяющее ребро $e_{j'_3}$, такое что, $i_3 < j'_3 \leqslant i_2$ то заменой e на $e_{j'_3}$ получаем граф $\hat{\Gamma}''$ с искомыми свойствами, иначе — найдем $n'_{i_4} \in \Gamma''_+$ и проведем с ней аналогичные рассуждения. В результате получится цепочка неравенств $f(e) \geqslant f(e_{j_3}) \geqslant f(e_{j_4}) \geqslant \ldots$. Поскольку j_m ограничены снизу и существует хотя бы одно соединяющее ребро, таковое будет найдено и заменой e на него будет построен искомый граф $\hat{\Gamma}''$, такой что $\mu(\hat{\Gamma}'') \leqslant \mu(\Gamma'')$ и $\chi_{n'_{k+1}}(\hat{\Gamma}'') = \chi_{n'_{k+1}}(\Gamma'') - 1$.

Таким образом цепочкой переходов $\Gamma'' \to \hat{\Gamma}'' \to \ldots$ можно придти к случаю $\tilde{\Gamma}''$, где $\chi_{n'_{k+1}}(\tilde{\Gamma}'') = 1$.

3.4 Заключение.

Предложенный алогритм может быть использован для частичной оптимизации схем широковещательных рассылок в метакластерах. В описанном методе никак не учитываются более или менее эффективные способы передачи данных в однородной коммуникационной среде. Однако, при непосредственной реализации алгоритма возможно учитывать различные подходы к оптимизации гомогенных широковещательных рассылок. Также не рассмотрена возможность параллельного исполнения некоторых операций, эта задача должна быть возложена на иные средства.