П.А. Павлов
Полесский государственный университет
г. Пинск, Республика Беларусь

ВРЕМЯ РЕАЛИЗАЦИИ РАСПРЕДЕЛЕННЫХ ПРОИЗВОДСТВЕННЫХ ПРОЦЕССОВ НА МНОЖЕСТВЕ КОНКУРИРУЮЩИХ ПОТОЧНЫХ ЛИНИЙ

Введение. Постоянное существование задач сверхвысокой сложности (проблемы экономики, экологии, космических исследований, изучения биологических и химических процессов, материаловедения и др.), которые нужно решать в указанных сферах, характеризуются большой размерностью, десятками сотен и миллионов независимых переменных и соответствующих ограничений. Подобные задачи актуальны и для Республики Беларусь. Это, прежде всего, задачи оптимизационного плана развития экономики страны или отдельного региона, задачи проектирования сложных систем (самолетов, энергетических котлов, ракетной техники), сооружений, технологических процессов, задачи эффективного использования спутников Земли для развития народного хозяйства, задачи

военного характера. Указанные задачи можно эффективно решать с помощью параллельных систем, применяя соответствующие математические модели и методы и используя идеи распараллеливания сложных процессов и обработки больших объемов данных и знаний.

1. Конструктивные элементы распределенных систем. Одним из режимов организации параллелизма является режим распределенной обработки, при котором используется множество интеллектуальных клиентов (обрабатывающих устройств, роботов, процессоров), достаточно удаленных друг от друга. Эффективная обработка данных при таком способе возможна только при низкой интенсивности потоков передач данных между обрабатывающими устройствами.

Основой для построения математических моделей параллельных систем, реализующих методы распределенной обработки решения задач, являются понятия процесса и реентерабельного ресурса.

Процесс будем рассматривать как последовательность наборов действий (блоков, команд) $I_s = (1, 2, ..., s)$. С целью наиболее эффективного решения задач синхронизации процессов, существенной минимизации накладных расходов и простоев обрабатывающих устройств, для процессов строятся расписания моментов запуска и окончания каждого из блоков. Моменты времени начала выполнения каждого блока определяются последовательностью $(t_1, t_2, ..., t_s)$. Предполагая, что блоки выполняются строго последовательно, в ходе своей реализации являются неделимыми и имеют длительности выполнения $d_j \geq 1$, $j = \overline{1,s}$, получим, что моменты времени завершения блоков определяются последовательностью вида $(t_1 + d_1, t_2 + d_2, ..., t_s + d_s)$.

Для ускорения выполнения процессы могут обрабатываться псевдопараллельно одним интеллектуальным клиентом или параллельно разными обрабатывающими устройствами, взаимодействуя между собой. Процессы, которые влияют на поведение друг друга путем обмена информацией, называют кооперативными или взаимодействующими процессами. В связи с этим, понятие процесса может быть использовано в качестве основного конструктивного элемента для построения распределенных систем в виде совокупности взаимодействующих процессов.

Понятие ресурса обычно используется для обозначения любых объектов распределенной системы, которые могут быть использованы процессом для своего выполнения. Для параллельных систем характерной является ситуация, когда одну и ту же последовательность команд или ее часть необходимо интеллектуальному клиенту выполнять многократно, которую будем называть реентерабельным ресурсом, а множество соответствующих процессов – конкурирующими. Решая проблему распределения многократно используемых ресурсов, как следствие неявно решаются задачи эффективного использования остальных категорий ресурсов. С этой точки зрения реентерабельный ресурс является ин-

тегрированным средством по запросам на выделяемые, повторно распределяемые и разделяемые ресурсы. С другой стороны, эффективно решая задачу распределения реентерабельных ресурсов, решаем проблему сокращения времени выполнения поставленных задач.

2. Концепция структурирования. Структурирование (декомпозиция) – это основной способ уменьшения времени использования больших реентерабельных ресурсов. Структурирование всегда предполагает то или иное разбиение многократно используемого ресурса на блоки с последующей организацией линейного порядка использования множества этих блоков. Разбиение на блоки многократно используемого ресурса – это наиболее важная и наиболее трудная область решений. Принципы структурирования реентерабельных ресурсов приобретают особенно фундаментальный характер в области создания эффективных распределенных систем. Основная идея метода структурирования состоит в обеспечении специального способа структурирования многократно используемого ресурса на блоки и организации одновременного использования этих блоков множеством процессов.

Пусть РР – реентерабельный ресурс, n, $n \geq 2$, число конкурирующих процессов за его использование. Требуется организовать процесс решения задачи таким образом, чтобы общее время выполнения n процессов, использующих РР, было минимальным.

Одной из стратегий решения данной задачи с p, $p \geq 2$, интеллектуальными клиентами является предоставление для каждого процесса отдельной копии РР. Но этот путь не всегда осуществим из–за ограниченного объема ресурсов и тем более трудно достижим в случае больших ресурсов, используемых в качестве многократно используемых. Поэтому при решении данной задачи применяется стратегия последовательного предоставления РР n процессам. В этом случае суммарное время выполнения n процессов составит величину $T_{sum} = nT$, где T – время выполнения каждым из процессов РР. Время T_{sum} можно существенно сократить, если обеспечить структурирование реентерабельного ресурса на блоки Q_1, Q_2, ..., Q_s с последующей конвейеризацией как блоков по процессам, так и процессов по интеллектуальным клиентам распределенной системы.

Структурирование РР на блоки осуществляется, как правило, либо исходя из физического смысла задачи на этапах создания математической модели и алгоритмов её решения, либо путём анализа многократно используемого ресурса с целью его декомпозиции. Число блоков, на которое осуществляется структурирование реентерабельного ресурса, зависит от количества процессов и обрабатывающих устройств, длительности выполнения ресурса, накладных расходов и других параметров.

Один из возможных механизмов взаимодействия процессов, интеллектуальных клиентов и блоков следующий. Блоки, процессы и клиенты нумеруются в порядке 1, 2, ..., s, 1, 2, ..., n и 1, 2, ..., p соответственно. Причем на множестве

блоков установлен линейный порядок их выполнения. Предполагается, что все n процессов используют одну и ту же копию структурированного многократно используемого ресурса. В дальнейшем под процессом будем понимать выполнение всех блоков в порядке 1, 2, ..., s. При этом процесс называется распределённым, если все блоки или часть из них выполняются на разных процессорах.

Специально выделенный организующий процесс предоставляет блоки структурированного РР Q_1, Q_2, ..., Q_s каждому из процессов в порядке 1, 2, ..., n. Если блок Q_j, $j = \overline{1,s}$, освобождается очередным i-м процессом, то он предоставляется (i + 1)–му процессу, а сам i-й процесс получает в своё распоряжение (j + 1)-й блок, либо переводится в состояние ожидания до освобождения (j+1)-го блока, $i = \overline{1,n-1}$, $j = \overline{1,s-1}$ и т. д. В случае распределённой обработки монополизация обрабатывающих устройств процессами не происходит, а блоки одного и того же процесса выполняются на разных устройствах.

Проблема обоснования метода структурирования многократно используемых ресурсов на параллельно выполняемые блоки носит комплексный характер, что порождает ряд сложных в математическом отношении проблем и задач. Для их решения требуется прежде всего построение математических моделей, адекватно отражающих различные аспекты взаимодействия множества процессов, с учетом их физической специфики, дополнительных накладных затрат и т. д. Анализ показывает, что на пути решения этой комплексной проблемы возникают математические задачи дискретно–комбинаторного характера, для решения которых эффективны аппарат теории графов, линейные диаграммы Ганта, теория расписаний, методы комбинаторной оптимизации.

3. Математическая модель параллельных распределенных производственных процессов. Будем говорить, что нижеперечисленные объекты образуют математическую модель конкурирующих поточных линий распределенного производственного процесса (реентерабельный ресурс): p, $p \geq 2$, специализированных рабочих мест (интеллектуальных клиентов), n, $n \geq 2$, поточных линий, s, $s \geq 2$, блоков структурированного производственного процесса, матрицу $T = [t_{ij}]$ $i = \overline{1,n}$, $j = \overline{1,s}$, времен выполнения блоков производственного процесса поточными линиями. Из физических соображений предполагается, что на множестве блоков производственного процесса установлен линейный порядок их выполнения 1,2, ..., s. Введем в рассмотрение параметр $\varepsilon > 0$, характеризующий дополнительные накладные расходы, затрачиваемые на организацию параллельного выполнения блоков производственного процесса множеством конкурирующих поточных линий.

Систему конкурирующих поточных линий будем называть неоднородной, если времена выполнения блоков распределенного производственного процесса Q_1, Q_2, ..., Q_s, разные на разных поточных линиях.

Система конкурирующих поточных линий называется однородной, если времена выполнения Q_j-го блока каждой поточной линией равны, т. е. $t_{ij} = t_j$, $i = \overline{1,n}$, $j = \overline{1,s}$.

Систему конкурирующих поточных линий будем называть одинаково распределенной, если времена выполнения всех блоков производственного процесса Q_j, $j = \overline{1,s}$, на каждой поточной линии совпадают и равны t_i для всех $i = \overline{1,n}$, т. е. справедлива цепочка равенств $t_{i1} = t_{i2} = \ldots = t_{is} = t_i$, для всех $i = \overline{1,n}$.

4. Режимы организации поточных линий, интеллектуальных клиентов и блоков структурированного производственного процесса. Взаимодействие поточных линий, специализированных рабочих мест и блоков структурированного производственного процесса подчинено следующим условиям:

- ни один из блоков производственного процесса не может обрабатываться одновременно более чем одним интеллектуальным клиентом;
- ни на одном из специализированных рабочих мест не может выполняться одновременно более одного блока;
- выполнение каждого блока производственного процесса осуществляется без прерываний;
- распределение блоков производственного процесса по рабочим местам для каждой из поточных линий осуществляется циклически по правилу: блок с номером $j = kp + i$, $j = \overline{1,s}$, $i = \overline{1,p}$, $k \geq 0$, распределяется на рабочее место с номером i.

Кроме того, введем дополнительные условия, которые определяют режимы взаимодействия поточных линий, обрабатывающих устройств и блоков производственного процесса:

- отсутствуют простои интеллектуальных клиентов при условии готовности блоков к выполнению, а также невыполнение блоков при наличии свободных специализированных рабочих мест;
- для каждой поточной линии момент завершения выполнения j-го блока на i-м рабочем месте совпадает с моментом начала выполнения следующего $(j + 1)$-го блока на $(i + 1)$-м рабочем месте, $i = \overline{1, p-1}$, $j = \overline{1, s-1}$;
- для каждого из блоков структурированного производственного процесса момент завершения его выполнения на l-й поточной линии совпадает с моментом начала его выполнения на $(l + 1)$-й на том же рабочем месте, $l = \overline{1,n-1}$.

Условия 1–5 определяют асинхронный режим взаимодействия поточных линий, обрабатывающих устройств и блоков распределенного производственного процесса, который предполагает отсутствие простоев обрабатывающих устройств, при условии готовности блоков, а также отсутствие невыполнений блоков при наличии свободных интеллектуальных клиентов. В этом режиме возможны ожидания, как блоков производственного процесса, так и освобождения обрабатывающих устройств.

Если к условиям 1–4 добавить поочередно условия 6 и 7 соответственно, то получим два синхронных режима:

– первый синхронный режим, определяемый условиями 1–4, 6, обеспечивает непрерывное выполнение блоков производственного процесса каждой поточной линией;

– второй синхронный режим, определяемый условиями 1–4, 7, обеспечивает непрерывное выполнение каждого блока всеми поточными линиями.

5. Время реализации распределенных производственных процессов на множестве конкурирующих поточных линий в асинхронном режиме. Интерес представляют задачи, связанные с получением математических соотношений, которые имеют как прямой, так и обратный характер. При постановке прямых задач условиями являются значения параметров системы конкурирующих поточных линий распределенного производственного процесса, а решением минимальные общие времена в различных режимах взаимодействия поточных линий, специализированных рабочих мест и блоков структурированного распределенного производственного процесса.

Обозначим минимальное общее время выполнения n неоднородных поточных линий распределенного производственного процесса на p специализированных рабочих местах в асинхронном режиме, с учетом введенного параметра ε, через $T_{n}^{ac}(p,n,s,\varepsilon)$. Для вычисления $T_{n}^{ac}(p,n,s,\varepsilon)$ рассмотрим случаи неограниченного ($2\le s\le p$) и ограниченного ($s>p$) параллелизма.

Пусть число блоков структурированного распределенного производственного процесса не превосходит числа интеллектуальных клиентов, т. е. ($2\le s\le p$). В этом случае без ограничения общности можно считать, что каждый Q_j-й блок закреплен за j-м обрабатывающим устройством, $j=\overline{1,s}$. Тогда для выполнения n поточных линий достаточно взять $p=s$ обрабатывающих устройств, а остальные $p-s$ будут не задействованы. Поэтому для определения минимального общего времени $T_{n}^{ac}(p,n,s,\varepsilon)$ можно воспользоваться функционалом задачи Беллмана–Джонсона, который в нашем случае будет иметь вид:

$$T_{n}^{ac}(s,n,s,\varepsilon)=\max_{1\le u_1\le u_2\le\ldots\le u_{s-1}\le n}\left[\sum_{i=1}^{u_1}t_{i1}^{\varepsilon}+\sum_{i=u_1}^{u_2}t_{i2}^{\varepsilon}+\ldots+\sum_{i=u_{s-1}}^{n}t_{is}^{\varepsilon}\right], \qquad (1)$$

где $t_{ij}^{\varepsilon}=t_{ij}+\varepsilon$, $i=\overline{1,n}$, $j=\overline{1,s}$, а $u_1, u_1, \ldots, u_{s-1}$ – целые числа.

В случае, когда $s=p$, функционал 1 будет иметь вид:

$$T_{n}^{ac}(p,n,p,\varepsilon)=\max_{1\le u_1\le u_2\le\ldots\le u_{p-1}\le n}\left[\sum_{i=1}^{u_1}t_{i1}^{\varepsilon}+\sum_{i=u_1}^{u_2}t_{i2}^{\varepsilon}+\ldots+\sum_{i=u_{p-1}}^{n}t_{is}^{\varepsilon}\right].$$

Задачу определения $T_{n}^{ac}(p,n,s,\varepsilon)$ можно решить более эффективно с помощью аппарата сетевых вершинно–взвешенных графов. В этом случае минимальное общее время в случае неограниченного параллелизма будет определяться длиной критического пути из начальной вершины в конечную.

В случае ограниченного параллелизма, т. е. когда $s > p$, $s = kp + r$, $k \geq 1$, $1 \leq r \leq p$ исходную матрицу времен выполнения блоков распределенного производственного процесса поточными линиями с учетом дополнительных системных расходов $T^{\epsilon} = [t_{ij}^{\epsilon}]$, $i = \overline{1,n}$, $j = \overline{1,kp + r}$, разбиваем на $(k + 1)$-у подматрицу $T_l^{\epsilon}, l = \overline{1,k + 1}$, размерностью $n \times p$ каждая. По каждой из подматриц строим линейную диаграмму Ганта, которые отражают во времени обработку очередных p блоков p обрабатывающими устройствами всех n поточных линий. Последняя диаграмма отражает выполнение последних r блоков производственного процесса (рис.1).

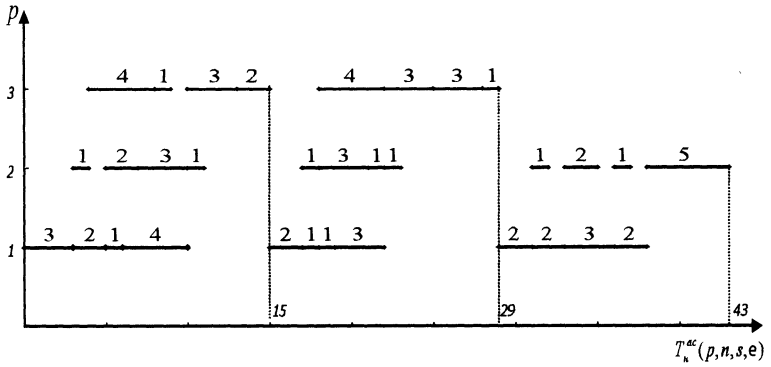

Рис.1 – Несовмещенная диаграмма Ганта

В этом случае общее время выполнения всех поточных линий будет определяться как сумма длин критических путей каждой из подряд идущих несовмещенных диаграммах Ганта. Это время можно существенно сократить, если воспользоваться приемом совмещения последовательных диаграмм Ганта по оси времени справа налево (рис.2).

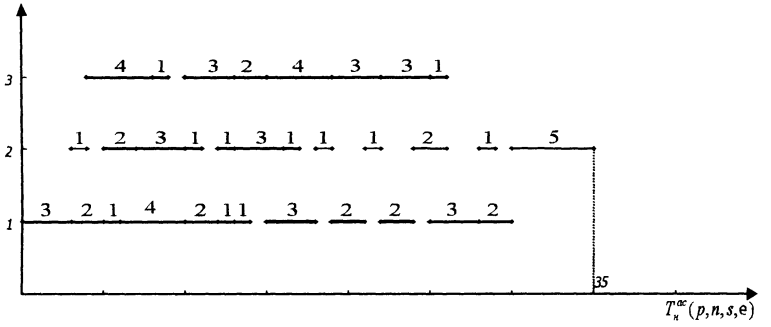

Рис.2 – Совмещенная диаграмма Ганта

Совмещенная диаграмма Ганта определяется результирующей матрицей T^* размерностью $(k+1)n\times(k+1)$p, которая является блочной, симметричной, верхнедиагональной относительно второй диагонали, типа Ганкелевой порядка $k+1$ (рис.3).

$$T^* = \begin{bmatrix} T_1^\varepsilon & T_2^\varepsilon & T_3^\varepsilon & \dots & T_k^\varepsilon & T_{k+1}^\varepsilon \\ T_2^\varepsilon & T_3^\varepsilon & T_4^\varepsilon & \dots & T_{k+1}^\varepsilon & 0 \\ T_3^\varepsilon & T_4^\varepsilon & T_5^\varepsilon & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ T_k^\varepsilon & T_{k+1}^\varepsilon & 0 & \dots & 0 & 0 \\ T_{k+1}^\varepsilon & 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Рис.3 – Матрица T^*

Если построить сетевой вершинно–взвешенный граф с весами, задаваемыми матрицей T^*, то минимальное общее время $T_u^{ac}(p,n,s,\varepsilon)$ выполнения неоднородных распределенных конкурирующих поточных линий распределенного производственного процесса в асинхронном режиме в случае ограниченного параллелизма определяется длиной критического пути из начальной вершины t_{11}^ε в конечную вершину $t_{(k+1)n,(k+1)p}^\varepsilon$ соответствующего сетевого вершинно–взвешенного графа, веса которого задаются матрицей T*.

Для однородной системы конкурирующих поточных линий, когда времена выполнения Q_j–го блока на каждой из поточных линий равны, т. е. $t_{ij}^\varepsilon = t_j^\varepsilon$, $i=\overline{1,n}$, $j=\overline{1,s}$, минимальное общее время в случае $2 \le s \le p$, в асинхронном режиме составляет величину $T_u^{ac}(p,n,s,\varepsilon)$ равную:

$$T_0^{ac}(p,n,s,\varepsilon) = T_\varepsilon^s + (n-1)\max_{1\le j\le s} t_j^s$$

где $(t_1^\varepsilon, t_2^\varepsilon, \dots, t_s^\varepsilon)$ – длительности выполнения каждого из блоков производственного процесса с учетом накладных расходов, а

$T_e^s = \sum_{j=1}^{s} t_j^s$ – длительность выполнения всего производственного процесса.

Для системы одинаково распределенных конкурирующих поточных линий, когда справедлива цепочка равенств $t_{i1}^\varepsilon = t_{i2}^\varepsilon = \dots = t_{is}^\varepsilon = t_i^\varepsilon$, $t_i^\varepsilon = t_i + \varepsilon$, для всех $i=\overline{1,n}$, минимальное общее время составляет величину $T_{op}^{ac}(p,n,s,\varepsilon)$, равную

$$T_{op}^{uc}(p,n,s,\varepsilon)=\begin{cases}T_e^n+(s-1)\max\limits_{1\le i\le n}t_i^e, & s\le p,\\ kT_e^n+(p-1)\max\limits_{1\le i\le n}t_i^e, & s=kp,\quad k>1,\\ (k+)T_e^n+(r-1)\max\limits_{1\le i\le n}t_i^e, & s=kp+r,\quad k\ge 1,1\le r<p,\end{cases}$$

где $T_e^n=\sum\limits_{i=1}^{n}t_i^e$ – суммарное время выполнения каждого из блоков Q_j, $j=\overline{1,s}$, всеми n конкурирующими поточными линиями.

В рамках математической модели организации конкурирующих поточных линий распределенного производственного процесса, исследуются первый и второй синхронные режимы. В условиях рассматриваемого режима для любых $p\ge 2$, $n\ge 2$, $s\ge 2$, $\varepsilon\ge 2$, получены временные характеристики реализации выполнения неоднородных, однородных и одинаково распределенных систем конкурирующих поточных линий.

Концепция построения распределенных производственных систем и их математическая основа может являться базой для эффективной организации технологических процессов в условиях ограниченности ресурсов, например, на предприятиях по производству медикаментов. В центре крупных городов создание поточных производств осложнено наличием свободных территорий, что, в свою очередь, влечет за собой создание ограниченного числа специализированных рабочих мест.

Заключение. Автором решены задачи определения минимального общего времени реализации множества распределенных кооперативных процессов, конкурирующих за многократно используемый ресурс. Решение было проведено на примере системы множества поточных линий распределенного производственного процесса в трех базовых режимах. При этом был использован аппарат теории расписаний и теории сетевых графов, что позволило разработать методы нахождения точных значений минимального общего времени выполнения распределенных производственных процессов в системе с ограниченным числом интеллектуальных клиентов с линейными оценками трудоемкости по числу блоков структурированного производственного процесса и поточных линий.