## Цифровая экономика и управление

УДК 004.75
DOI:10.25729/ESI.2024.35.3.009

# Сравнительный анализ систем управления научными рабочими процессами

**Воскобойников Михаил Леонтьевич, Феоктистов Александр Геннадьевич**

Институт динамики систем и теории управления им. В.М. Матросова СО РАН,
Россия, Иркутск, *mikev1988@mail.ru*

**Аннотация.** Стремительное развитие параллельных и распределенных вычислительных систем, телекоммуникационных технологий и облачных платформ обеспечило возможность разработки и применения научных приложений с целью подготовки и проведения крупномасштабных экспериментов с большими массивами данных. Зачастую создаваемые приложения предполагают сложную схему решения задач, базирующуюся на интегрированном выполнении процессов передачи, обработки и анализа данных, ресурсоемких вычислений и принятия решений. При этом математическое и программное обеспечение приложений может реализовываться различными группами специалистов из разных организаций и ориентироваться на разнородные вычислительные ресурсы. Все это обусловливает необходимость использования развитых средств проектирования, реализации, развертывания и выполнения научных рабочих процессов в рамках единой распределенной вычислительной среды, интегрирующей в конечном счете алгоритмические знания, программно-аппаратное обеспечение, данные и разнообразные сервисы. В настоящее время в качестве таких средств, как правило, выступают системы управления рабочими процессами. В этой связи статья посвящена обсуждению текущего состояния известных систем управления рабочими процессами, а также рассмотрению проблем, связанных с научными рабочими процессами для различных вычислительных сред. Актуализируются проблемы разработки и применения подобных систем, которые в настоящее время не решены в полной мере. В частности, отмечается необходимость учета предметной специфики и обеспечения масштабирования вычислений, востребованность сервис-ориентированных приложений, эффективность эксплуатации гетерогенных сред, интегрирующих высокопроизводительные ресурсы пользователей, кластерные ресурсы центров коллективного пользования, Grid-систем и облачных платформ.

**Ключевые слова:** распределенные вычисления, научные рабочие процессы, системы управления рабочими процессами

**Цитирование:** Воскобойников М.Л. Сравнительный анализ систем управления научными рабочими процессами / М.Л. Воскобойников, А.Г. Феоктистов // Информационные и математические технологии в науке и управлении, 2024. – № 3(35). – С.102-111. – DOI: 10.25729/ESI.2024.35.3.009.

**Введение.** В настоящее время решение крупномасштабных фундаментальных и прикладных задач зачастую осуществляется на основе распределенных научных приложений (РНП), характеризующихся модульной структурой их прикладного программного обеспечения (ПО), представленной в виде вычислительной модели, развитым системным ПО (совокупностью программ, обеспечивающих построение приложения, организацию вычислений с помощью его прикладных модулей, обработку данных и взаимодействие различных категорий пользователей с приложением) и ориентацией на решение определенного класса задач. В современных приложениях схема решения задачи реализуется научным рабочим процессом (НРП, англ., Scientific Workflow) – информационно-вычислительной структурой, отражающей бизнес-логику предметной области относительно данных, ПО (набора прикладных модулей) и вычислительных ресурсов в процессе проведения экспериментов.

Системы создания приложений на основе НРП относятся к классу систем управления рабочими процессами (англ., Workflow Management Systems – WMS). В общем случае WMS предоставляют программную инфраструктуру для настройки, выполнения и мониторинга определенной последовательности задач, организованных в виде рабочего процесса со сложными зависимостями между решаемыми задачами [1].

В процессе разработки НРП возникает необходимость решения следующих важных проблем: согласованного использования разнородных вычислительных ресурсов; учета специфики предметных областей решаемых задач; интеграции системного и прикладного ПО при подготовке и проведении экспериментов; стандартизации спецификаций объектов предметной области и вычислительной модели, форматов представления и протоколов передачи данных, доступа пользователей к подсистемам приложения; управления вычислениями в интегрированной среде посредством взаимодействия исполнительной подсистемы с системами мониторинга, локальными менеджерами ресурсов (ЛМР, англ., Local Resource Manager – LRM) в узлах среды, и метапланировщиками, работающими на уровне среды в целом; обеспечения масштабирования вычислений; поддержки технологии In-Memory Data Grid (IMDG) с целью ускорения обработки данных [2].

Необходимость учета предметной специфики и обеспечения масштабирования вычислений обусловливает переход от использования вычислительной среды общего назначения (Grid-системы, облачной платформы, ресурсов центров коллективного пользования и др.) к предметно-ориентированной вычислительной среде (ПОВС) с рациональным сочетанием возможностей включаемых в ее состав информационно-вычислительных ресурсов, потребностей и накладных затрат, обусловливаемых особенностями предметных областей для конкретных классов решаемых задач. При этом особое внимание научного сообщества уделяется развитию сервис-ориентированного подхода к организации и применению НРП.

Статья посвящена сравнительному анализу функциональных возможностей известных WMS относительно процессов разработки НРП в различных распределенных вычислительных средах.

**Среда выполнения приложений.** В общем случае рассматривается гетерогенная распределенная вычислительная среда (ГРВС), которая может интегрировать собственные кластерные ресурсы пользователей, а также кластерные ресурсы ЦКП, Grid и облачные платформы. Кластерные ресурсы для выполнения приложений могут объединяться и управляться на основе виртуализации или контейнеризации.

РНП [3] базируется на традиционном понятии пакета прикладных программ [4] и представляет собой комплекс взаимосвязанных прикладных программ и средств системного обеспечения (программных и языковых), предназначенный для автоматизации решения определенного класса задач в конкретной предметной области в ГРВС. В то же время РНП характеризуется следующими особенностями:

- ориентация на решение класса задач, требующих проведения расчетов с задействованием больших объемов вычислительных ресурсов (процессорного времени, оперативной памяти, дискового пространства и других характеристик);
- общая задача подразумевает параллельное решение множества ее слабо связанных или полностью независимых подзадач;
- алгоритмические знания (прикладное ПО) представляются в виде набора программных модулей, реализующих схемы решения задач;
- в процессе вычислений варианты данных параллельно обсчитываются экземплярами модулей;
- не предполагается интенсивного взаимодействия между параллельными вычислительными процессами;
- НРП представляет собой схему решения задачи;
- вычисления выполняются, как правило, по одной слабо меняющейся схеме, требующей динамического управления процессом вычислений;
- управление вычислениями осуществляет диспетчер НРП;

– необходимо использование связующего ПО.

Общая схема функционирования ГРВС для выполнения РППП приведена на рис. 1.

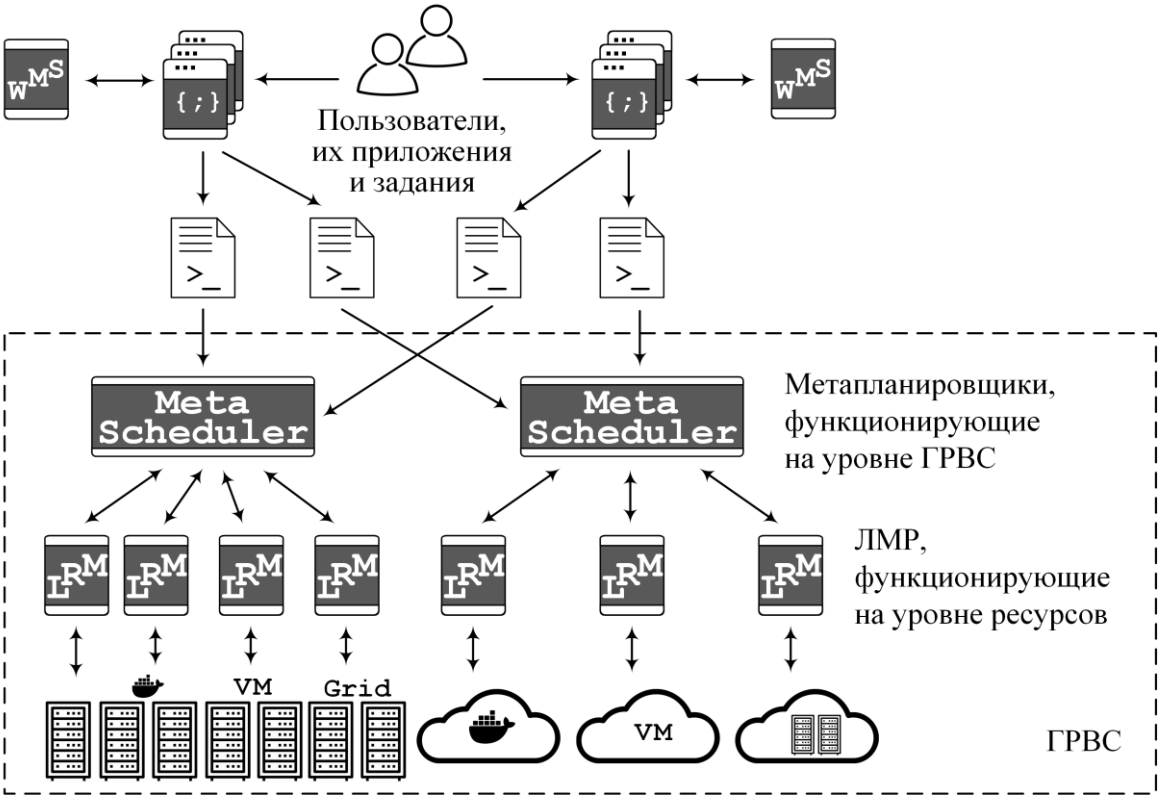


**Рис. 1.** Общая схема функционирования ГРВС

Пользователи приложений формируют задания в виде текстовых спецификаций вычислительных процессов, определяемых НРП. Такие спецификации включают сведения о требуемых ресурсах, исполняемом прикладном ПО, входных и выходных данных, а также иную необходимую дополнительную информацию. Например, дополнительная информация может определять критерии качества решения своей задачи: время, стоимость, надежность, безопасность и другие характеристики.

Основными компонентами ГРВС являются высокопроизводительные кластеры, кластеры виртуализированных и контейнеризированных ресурсов, которые используются в рамках таких инфраструктур, как ЦКП, кластерные Grid и облачные платформы. ГРВС может интегрировать компоненты перечисленных выше инфраструктур.

В общем случае управление вычислениями в ГРВС включает следующие уровни:

– уровень приложений, где выполняется планирование вычислений, построение НРП и выбор вычислительных ресурсов для его выполнения;
– уровень среды, на котором потоки заданий попадают в очередь метапланировщиков ГРВС, поддерживающих взаимодействие с выбранными ресурсами, и затем, в соответствии с дисциплинами диспетчеризации, им назначаются конкретные ресурсы;
– уровень ресурсов, где задания распределяются ЛМР между узлами с целью их дальнейшего выполнения.

**Системы управления научными рабочими процессами.** НРП представляется специализированной формой графа, которая описывает вычислительные процессы, их зависимости в рамках сбора, обработки и анализа данных. Он представляет собой бизнес-логику предметной области в применении предметно-ориентированных данных и ПО (набора прикладных модулей) для решения задач в этой предметной области.

Различают два типа НРП: абстрактный и конкретный. Абстрактный НРП описывается в виде абстрактной схемы решения задачи без обращения к конкретным ресурсам. Это позволяет определять НРП без детальной конкретизации его реализации на низком программно-аппаратном уровне. Операции абстрактного НРП могут реализовываться различным прикладным ПО и отображаться на любые подходящие ресурсы с помощью механизмов планирования вычислений и управления ресурсами. В случае конкретного НРП его операции связаны с предопределенным ПО, выполняемом на заданных ресурсах.

Построение НРП осуществляется по процедурной или непроцедурной постановкам задач. В первом случае задается последовательность операций НРП и затем осуществляется информационное планирование его входных и выходных параметров. Во втором случае, задача формулируется следующим образом: «вычислить значения целевых (выходных) параметров по заданным значениям исходных (входных) параметров». Далее на вычислительной модели (описании предметной области – ее параметров, вычислительных операций и операций обработки данных, а также отношений между параметрами и операциями) осуществляется автоматическое планирование последовательности операций НРП, необходимых для решения поставленной задачи. Как правило, планирование вычислений осуществляется одним из системных компонентов WMS. Непроцедурная постановка задачи позволяет разработчикам приложения и его пользователям работать на более высоком уровне абстракции. Они скрывают детали реализации и позволяют сосредоточиться на описании задачи и отношений между объектами.

WMS – это программное средство, частично автоматизирующее определение и выполнение НРП, а также управление им в соответствии с его информационно-вычислительной структурой. Для задания НРП часто используется ориентированный ациклический граф (DAG). В общем случае вершины и ребра DAG соответственно представляют прикладные программные модули и потоки данных между ними.

Следует отметить следующие WMS, которые хорошо известны, поддерживаются, развиваются и широко применяются на практике: Uniform Interface to Computing Resources (UNICORE) [5], Directed Acyclic Graph Manager (DAGMan) [6], Pegasus [7], Apache Airflow (AA) [8], HyperFlow [9], Workflow-as-a-Service Cloud Platform (WaaS CP) [10], Galaxy [11] и Orlando Tools (OT) [12]. UNICORE, DAGMan, Pegasus, AA и OT находятся на лидирующих позициях в области традиционного управления НРП. Они основаны на модульном подходе к созданию таких процессов. В рамках НРП общая задача разбивается на набор задач, обычно представленных в виде DAG. Как правило, UNICORE, DAGMan, Pegasus, AA, OT и другие подобные системы предоставляют свою собственную нестандартизированную модель НРП. Это затрудняет совместное использование спецификаций рабочих процессов и ограничивает совместимость приложений, разработанных с использованием различных WMS.

Комплексная поддержка веб-сервисов является актуальным направлением современных WMS. Использование веб-сервисов значительно расширяет вычислительные возможности приложений, основанных на НРП. WMS, ориентированные на работу с веб-сервисами, позволяют гибко и динамично интегрировать различные НРП от разных разработчиков за счет управления данными и вычислениями при выполнении этих процессов. В этом направлении успешно развиваются HyperFlow, WaaS CP и Galaxy.

В научных приложениях для решения различных классов задач в области мониторинга окружающей среды, особое значение имеет интеграция с геоинформационными системами посредством специализированных сервисов веб-обработки (WPS). Примерами проектов, направленных на создание и применение таких программных комплексов, являются GeoJModelBuilder (GJMB) [13] и Business Process Execution Language (BPEL) Designer Project

(DP) [14]. GJMB и BPEL DP относятся к классу WMS. GJMB – это фреймворк для управления и координации геопространственных датчиков, данных и прикладного ПО в среде, основанной на НРП. BPEL DP реализует интеграцию веб-сервисов и WPS с помощью рабочих процессов на основе BPEL. BPEL определяет модель для описания поведения сервис-ориентированных НРП в терминах взаимодействий (совокупности сообщений) между процессами и их партнерами (внешними сервисами). Другими преимуществами BPEL являются следующие:

- НРП могут не только вызывать веб-сервисы, но и сами быть представленными в виде сервисов;
- широкий спектр элементов управления и работы с данными, включающий элементы определения сложных структур данных и параллельных процессов их обработки, циклы, ветвления, подпроцессы, элементы реализации асинхронного взаимодействия веб-сервисов и др.;
- использование языка Web Service Description Language (WSDL) для описания интерфейсов веб-сервисов обеспечивает гибкую интеграцию с другими НРП и веб-приложениями;
- детальное описание НРП реализует оркестровку внутренних и внешних сущностей процесса, а спецификация процесса обмена сообщениями отражает хореографию внешних сущностей (вызываемых веб-сервисов).

Ключевые возможности WMS, которые наиболее важны с точки зрения разработки РНП на основе НРП, приведены в таблице 1.

**Таблица 1.** Возможности WMS для поддержки НРП

| WMS | Характеристики процесса построения НРП | | | | | | | Особенности организации вычислений | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ | $c_8$ | $c_9$ | $c_{10}$ | $c_{11}$ | $c_{12}$ | $c_{13}$ |
| UNICORE | + / - / - / - /- / - | + /+ | + | - / + | + / - | + | - | + / + / - | - / + / - | - | - | - | - |
| DAGMan | - / + / - / - /- / - | - /- | + | + / - | + / - | - | - | + / + / - | + / + / + | + | - | - | - |
| Pegasus | + / - / - / - /- / - | - /- | + | + / - | + / - | - | - | + / + / - | + / + / + | + | - | - | - |
| AA | - / - / + / - /- / - | - /- | - | - / + | + / - | - | - | + / + / + | - / - / + | + | - | - | - |
| HyperFlow | - / - / - / - /+ / + | - /- | + | + / + | + / - | + | - | + / - / - | - / - / + | - | - | - | - |
| WaaS CP | + / - / - / - /- / - | - /- | - | + / - | + / - | + | - | + / - / - | - / + / + | + | - | - | - |
| Galaxy | + / - / - / - /- / - | - / + | + | + / + | + / - | + | - | + / - / - | - / - / + | - | - | - | - |
| OT | + / - / - / - /- / - | + / + | + | + / + | + / + | - | - | + / + / + | + / + / + | - | + | + | - |
| GJMB | + / - / - / - /- / - | - / - | + | + / + | + / - | + | + | + / - / - | + / + / + | - | - | - | - |
| BPEL DP | - / - / - / + /- / - | + /+ | - | + / + | + / - | + | + | + / - / - | + / + / + | - | - | - | - |

Рассматриваются следующие возможности WMS:

- $c_1$ – поддержка языков описания НРП: XML-like / Script-like / Python / BPEL / JavaScript / JSON;
- $c_2$ – поддержка управляющих языковых конструкций: ветвления / циклы;
- $c_3$ – поддержка возможности включения в НРП системных операций;
- $c_4$ – поддержка типов НРП: абстрактный / конкретный;
- $c_5$ – поддержка постановок задач: процедурная / непроцедурная;
- $c_6$ – поддержка вызова веб-сервисов в НРП;
- $c_7$ – поддержка стандарта WPS-сервисов;
- $c_8$ – уровень поддержки распараллеливания вычислений: задача / данные / конвейер;
- $c_9$ – поддержка следующих типов вычислительной среды: кластер / Grid / облачная среда;
- $c_{10}$ – необходимость подключения дополнительного связующего ПО;
- $c_{11}$ – поддержка технологии IMDG;

- $c_{12}$ – автоматизация непрерывной интеграции ПО;
- $c_{13}$ – автоматизация контейнеризации ПО.

WMS используют XML-подобные или скриптовые языки для описания предметных областей приложений и самих НРП. BPEL характеризуется рядом преимуществ для спецификации НРП, перечисленных выше. Однако его языковых средств зачастую бывает недостаточно, чтобы описать сложную вычислительную модель. Поэтому целесообразно применять комбинацию XML-подобного языка и BPEL. Ветвления и циклы в рабочем процессе позволяют при необходимости более гибко организовать вычисления. UNICORE, OT и BPEL DP предоставляют такую поддержку в полном объеме для построения НРП, не связанного с DAG. Дополнительным преимуществом UNICORE, DAGMan, Pegasus, HyperFlow, Galaxy, GJMB и OT является возможность включения системных операций в структуру рабочего процесса. К таким операциям относятся пред- и постобработка данных, мониторинг среды, взаимодействие с компонентами локальных менеджеров ресурсов и т.д.

HyperFlow, WaaSCP, Galaxy, OT, GJMB и BPEL DP позволяют строить как абстрактные, так и конкретные НРП. При этом OT обеспечивает оба вида постановок задач, на основе которых выполняется построение НРП: процедурную и непроцедурную.

HyperFlow, WaaSCP, Galaxy, GJMB и BPEL DP ориентированы на работу с веб-сервисами. Кроме того, GJMB и BPEL DP поддерживают стандарт WPS.

Большая часть рассматриваемых WMS являются системами общего назначения и могут успешно применяться в различных предметных областях. Некоторые системы являются более специализированными. Например, Galaxy ориентирована на поддержку проведения геномных исследований, а GJMB и BPEL DP позволяют решать задачи в области геоинформатики.

Уровень задач означает, что задачи, определенные структурой рабочего процесса, выполняются на независимых узлах. На уровне данных набор данных делится на подмножества. Каждое подмножество обрабатывается на отдельном узле всеми или частью прикладных программных модулей, включенных в рабочий процесс. На уровне конвейера последовательное выполнение прикладных модулей для обработки данных выполняется одновременно на разных подмножествах набора данных. AA и OT предоставляют все уровни распараллеливания вычислений. DAGMan, Pegasus, OT, GJMB и BPEL DP допускают использование всех типов вычислительной среды. Однако системе Pegasus требуется интеграция с DAGMan для управления заданиями приложений на уровне среды. OT предоставляет системные операции для поддержки технологии IMDG на основе связующего ПО Apache Ignite. Инструменты автоматизации непрерывной интеграции прикладного и системного ПО предоставляются в OT. В то же время обеспечение автоматизации процессов контейнеризации ПО является не решенной в полной мере проблемой для всех известных WMS. Практически во всех WMS слабо проработаны вопросы моделирования вычислительных процессов. Решение упомянутых вопросов требует включения в WMS дополнительных средств взаимодействия с контрольно-измерительными системами, профилировщиками ПО, инструментами анализа данных и моделирования работы как отдельных информационно-вычислительных компонентов ГРВС, так и всей среды в целом.

К сожалению, из-за разницы в функциональных и системных возможностях WMS затруднительно провести адекватное сравнение трудозатрат их пользователей на разработку приложений, подготовку и проведение вычислительных экспериментов, а также оценить эффективность использования вычислительных ресурсов при решении практических задач. Однако некоторые результаты сравнительного анализа управления выполнением ряда известных рабочих процессов метапланировщиками на уровне среды можно найти в [15].

**Направления развития WMS.** Вопросы разработки и применения НРП в разных областях исследований требуют постоянного их изучения, сравнения и развития с целью поддержки междисциплинарного сотрудничества. Потребность совместного использования НРП, миграции этих рабочих процессов между различными вычислительными средами, их адаптации к различным категориям пользователей, языкам и программным средствам обоснованно обусловливают необходимость совершенствования WMS. На основе проведенного сравнительного анализа можно определить направления развития функциональных возможностей WMS. В том числе выделить следующие важные аспекты дополнительных разработок для таких систем:

- предоставление дружественного пользовательского интерфейса для взаимодействия с компонентами инструментального комплекса и создаваемых сервис-ориентированных приложений разным категориям пользователей (разработчикам, администраторам и пользователям), поддерживающего как текстовые языки представления структурированных данных, так и языки для ввода информации с помощью веб-форм с последующим автоматическим конвертированием полученных спецификаций на соответствующие текстовые языки;
- корректное описание вычислительной модели предметной области (параметров, операций, логических выражений, продукций и модулей, вычислительных ресурсов, элементов вычислительной истории, характеристик административных политик управления ресурсами и др., а также отношений между перечисленными объектами);
- работа со сложными структурами данных, такими, как составные параметры и параллельные списки данных, а также неструктурированными файловыми данными;
- статическое планирование вычислений по непроцедурной постановке задачи, построение НРП по процедурной постановке задачи с использованием различных операторов ветвления и циклов, а также статико-динамическое планирование расчетных схем и распределение возникающей нагрузки на ресурсы при работе с параллельными списками данных;
- обеспечение средств создания и применения сервис-ориентированных НРП при сохранении поддержки традиционных НРП с исполняемыми модулями;
- возможность включения в состав НРП системных операторов обработки и анализа данных, статико-динамического планирования, конвейеризации вычислений, параллельного выполнения алгоритмов;
- генерация НРП на языке программирования общего назначения с целью их последующего автономного запуска;
- автоматизация непрерывной интеграции и контейнеризации ПО, включая моделирование вычислительных процессов и тестирование ПО на ресурсах ГРВС.

**Заключение.** Разработка научных приложений, основанных на рабочих процессах, является перспективной парадигмой для решения широкого спектра фундаментальных и прикладных задач в различных областях знаний. Использование высокопроизводительных вычислительных систем обеспечивает высокую эффективность процесса решения. В этой связи научное сообщество прилагает значительные усилия в развитии данной парадигмы. В частности, пристальное внимание уделяется системам управления НРП.

В рамках проведенного исследования рассмотрены основные понятия, связанные с НРП и системами управления ими. Проведен сравнительный анализ известных систем управления НРП. Оценены тенденции современного развития РНП на основе сервис-ориентированных НРП. Обсуждены вопросы и проблемы исследований в этой области, не решенные в полной мере, в том числе определены актуальные направления развития систем управления НРП.

Дальнейшие исследования связаны с разработкой нового инструментального комплекса, развивающего функциональные возможности ОТ применительно к сервис-ориентированным приложениям, включая представление НРП на языке BPEL и поддержку работы с WPS-сервисами, а также автоматизацию процессов генерации программ на языках программирования общего назначения на основе НРП и контейнеризации прикладного и системного ПО в ГРВС. Реализация перечисленных выше средств позволит успешно использовать инструментальный комплекс для построения цифровых двойников компонентов информационно-вычислительных систем и их программно-аппаратных инфраструктур на основе системных НРП, дополнительно направленных на анализ и моделирование характеристик, свойств, состояний и процессов вычислительной среды.

**Благодарности.** Исследование проведено в рамках проекта № FWEW-2021-0005 «Технологии разработки и анализа предметно-ориентированных интеллектуальных систем группового управления в недетерминированных распределенных средах» (рег. № 121032400051-9) при поддержке Министерства науки и высшего образования РФ.

## Список источников

1. Schael T. Workflow management systems for process organisations. Lecture notes in computer science, 1996, vol. 1096, p. 208, DOI:10.1007/978-3-662-21574-6.
2. Zhang H., Chen G., Ooi B.C. In-memory big data management and processing: A survey. IEEE Transactions on knowledge and data engineering, 2015, vol. 27, no. 7, pp. 1920–1948, DOI:10.1109/TKDE.2015.2427795.
3. Феоктистов А.Г. Автоматизация разработки и применения распределенных пакетов прикладных программ / А.Г. Феоктистов, И.А. Сидоров, С.А. Горский // Проблемы информатики, 2017. – № 4. – С. 61-78.
4. Горбунов–Посадов М.М. Системное обеспечение пакетов прикладных программ / М.М. Горбунов–Посадов, Д.А. Корягин, В.В. Мартынюк. – М.: Наука, 1990. – 208 с.
5. UNICORE. Available at: https://www.unicore.eu/ (accessed: 09.04.2024).
6. DAGMan. Available at: https://htcondor.org/dagman/dagman.html (accessed: 04/09/2024).
7. Deelman E., Singh G., Su M.H. et al. Pegasus: A framework for mapping complex scientific workflows onto distributed systems. Scientific programming, 2005, vol. 13, pp. 219–237, DOI:10.1155/2005/128026.
8. Singh P. Learn PySpark: Build Python-based machine learning and deep learning models. Apress, 2019, pp. 67–84.
9. Balis B. HyperFlow: A model of computation, programming approach and enactment engine for complex distributed workflows. Future generation computer systems, 2016, vol. 55, pp. 147–162. DOI:10.1016/j.future.2015.08.015.
10. Hilman M.H., Rodriguez M.A., Buyya R. Workflow-as-a-service cloud platform and deployment of bioinformatics workflow applications. Knowledge management in the development of data-intensive systems, 2021, pp. 205–226, DOI:10.1201/9781003001188-14.
11. Jalili V., Afgan E., Gu Q. et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses. Nucleic acids research, 2020, vol. 48, no. W1, pp. W395-W402, DOI:10.1093/nar/gkaa434.
12. Gorsky S.A. Continuous integration, delivery, and deployment for scientific workflows in Orlando Tools. Proceedings of the 2nd international workshop on information, computation, and control systems for distributed environments, CEUR-WS Proceedings, 2020, vol. 2638, pp. 118-128, DOI:10.47350/ICCS-DE.2020.11.
13. Yue P., Zhang M., Tan Z. A geoprocessing workflow system for environmental monitoring and integrated modelling. Environmental modelling and software, 2015, vol. 69, pp. 128–140, DOI:10.1016/j.envsoft.2015.03.017.
14. Tan X., Jiao J., Chen N. et al. Geoscience model service integrated workflow for rain-storm waterlogging analysis. International journal of digital Earth, 2021, vol. 14, pp. 851–873, DOI:10.1080/17538947.2021.1898686.
15. Feoktistov A. Tender of computational works in heterogeneous distributed environment. CEUR-WS Proceedings, 2020, vol. 2638, pp. 99–108, DOI:10.47350/ICCS-DE.2020.09.

***Воскобойников Михаил Леонтьевич.*** *Младший научный сотрудник института динамики систем и теории управления им. В.М. Матросова СО РАН. Область научных интересов – распределенные вычисления, системы управления рабочими процессами, автоматизация разработки сервис-ориентированных приложений. AuthorID: 1102663, SPIN: 3417-0258, ORCID: 0000-0003-3034-4907, mikev1988@mail.ru, Россия, г. Иркутск, Лермонтова,134.*

***Феоктистов Александр Геннадьевич.*** *Д.т.н., доцент, главный научный сотрудник института динамики систем и теории управления им. В.М. Матросова СО РАН. Область научных интересов – распределенные вычисления, системы управления рабочими процессами, автоматизация разработки научных приложений, концептуальное моделирование. AuthorID: 1535, SPIN: 5743-1777, ORCID: 0000-0002-9127-6162, agf65@yandex.ru, Россия, г. Иркутск, Лермонтова,134.*

UDC 004.75
DOI:10.25729/ESI.2024.35.3.009

# Comparative analysis of scientific workflow management systems

**Mikhail L. Voskoboinikov, Alexander G. Feoktistov**

Matrosov Institute for System Dynamics and Control Theory of SB RAS,
Russia, Irkutsk, *mikev1988@mail.ru*

**Abstract.** The rapid evolution of parallel and distributed computing systems, telecommunication technologies, and cloud platforms has enabled the development and use of scientific applications to prepare and conduct large-scale experiments with large amounts of data. Often, the applications implement a complex problem-solving scheme based on the integrated execution of processes for data transfer, processing and analysis, resource-intensive computation, and decision-making. At the same time, the mathematical models and software of applications may be developed by different groups of specialists from different organizations and focused on heterogeneous computing resources. This requires the use of advanced tools for the design, implementation, deployment, and execution of scientific workflows within a single distributed computing environment, ultimately integrating algorithmic knowledge, software and hardware used, data, and various services. Today, such tools are usually workflow management systems. In this context, the paper is dedicated to discuss the current state of known workflow management systems, as well as to address the problems associated with the development and use of scientific workflows in different computing environments. The problems associated with the development and use of such systems, which are currently not fully solved, are highlighted. In particular, we point out the need to take into account subject domain specificities, the computation scaling, the demand for service-oriented applications, and the efficiency of using heterogeneous distributed environments that integrate high-performance user resources, cluster resources of shared use centers, Grid systems, and cloud platforms.

**Keywords:** distributed computing, scientific workflows, workflow management systems

**Acknowledgements:** The study was supported by the Ministry of Science and Higher Education of the Russian Federation, project no. № FWEW-2021-0005 “Technologies for the development and analysis of subject-oriented intelligent group control systems in non-deterministic distributed environments”.

## References

1. Schael T. Workflow management systems for process organisations. Lecture notes in computer science, 1996, vol. 1096, p. 208, DOI:10.1007/978-3-662-21574-6.
2. Zhang H., Chen G., Ooi B.C. In-memory big data management and processing: A survey. IEEE Transactions on knowledge and data engineering, 2015, vol. 27, no. 7, pp. 1920–1948, DOI:10.1109/TKDE.2015.2427795.
3. Feoktistov A.G., Sidorov I.F., Gorsky S.A. Avtomatizaciia razrabotki i primeneniia raspredelennykh paketov prikladnykh programm [Automation of development and application of distributed applied software packages]. Problemy informatiki [Problems of Informatics], 2017, no. 4, pp. 61-78.
4. Gorbunov–Posadov M. M., Koryagin D.A., Martynyuk V.V. Sistemnoe obespechenie paketov prikladnyh programm [System support of applied software packages]. Moskow, Nauka, 1991, 208 p.
5. UNICORE. Available at: https://www.unicore.eu/ (accessed: 09.04.2024).
6. DAGMan. Available at: https://htcondor.org/dagman/dagman.html (accessed: 04/09/2024).
7. Deelman E., Singh G., Su M.H. et al. Pegasus: A framework for mapping complex scientific workflows onto distributed systems. Scientific programming, 2005, vol. 13, pp. 219–237, DOI:10.1155/2005/128026.
8. Singh P. Learn PySpark: Build Python-based machine learning and deep learning models. Apress, 2019, pp. 67–84.

9. Balis B. HyperFlow: A model of computation, programming approach and enactment engine for complex distributed workflows. Future generation computer systems, 2016, vol. 55, pp. 147–162. DOI:10.1016/j.future.2015.08.015.
10. Hilman M.H., Rodriguez M.A., Buyya R. Workflow-as-a-service cloud platform and deployment of bioinformatics workflow applications. Knowledge management in the development of data-intensive systems, 2021, pp. 205–226, DOI:10.1201/9781003001188-14.
11. Jalili V., Afgan E., Gu Q. et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses. Nucleic acids research, 2020, vol. 48, no. W1, pp. W395-W402, DOI:10.1093/nar/gkaa434.
12. Gorsky S.A. Continuous integration, delivery, and deployment for scientific workflows in Orlando Tools. Proceedings of the 2nd international workshop on information, computation, and control systems for distributed environments, CEUR-WS Proceedings, 2020, vol. 2638, pp. 118-128, DOI:10.47350/ICCS-DE.2020.11.
13. Yue P., Zhang M., Tan Z. A geoprocessing workflow system for environmental monitoring and integrated modelling. Environmental modelling and software, 2015, vol. 69, pp. 128–140, DOI:10.1016/j.envsoft.2015.03.017.
14. Tan X., Jiao J., Chen N. et al. Geoscience model service integrated workflow for rain-storm waterlogging analysis. International journal of digital Earth, 2021, vol. 14, pp. 851–873, DOI:10.1080/17538947.2021.1898686.
15. Feoktistov A. Tender of computational works in heterogeneous distributed environment. CEUR-WS Proceedings, 2020, vol. 2638, pp. 99–108, DOI:10.47350/ICCS-DE.2020.09.

***Voskoboinikov Mikhail Leontevich.*** *Junior Researcher at the Matrosov institute for system dynamics and control theory of SB RAS. The main direction of research – distributed computing, workflow management systems, automation of developing service-oriented applications. AuthorID: 1102663, SPIN: 3417-0258, ORCID: 0000-0003-3034-4907, mikev1988@mail.ru, Russia, Irkutsk, Lermontova, 134.*

***Feoktistov Alexander Gennadevich.*** *Doctor of technical sciences, associate professor, chief researcher at the Matrosov institute for system dynamics and control theory of SB RAS. The main direction of research – distributed computing, workflow management systems, automation of developing scientific applications, conceptual modeling. AuthorID: 1535, SPIN: 5743-1777, ORCID: 0000-0002-9127-6162, agf65@yandex.ru, Russia, Irkutsk, Lermontova, 134.*

*Статья поступила в редакцию 10.04.2024; одобрена после рецензирования 19.09.2024; принята к публикации 08.10.2024.*

*The article was submitted 04/10/2024; approved after reviewing 09/19/2024; accepted for publication 10/08/2024.*