

Неупокоев М.О., Булах Б.В.

Институт прикладного системного анализа НТУУ "КПИ", Киев, Украина

Workflow-системы для решения научных задач

Научные workflow-системы – это специализированные системы управления потоками работ, разработанные специально для выполнения комплекса вычислительных операций, необходимых для решения научных задач [1]. Растущий интерес к научным workflow-системам совпал с ростом интереса к e-Science технологиям и приложениям, и с развитием Грид-компьютинга. Видение e-Science заключается в том, что ученые, невзирая на разделяющие их расстояния, в состоянии сотрудничать при проведении крупномасштабных научных экспериментов и обмениваться новыми знаниями с использованием распределенных вычислительных систем. Научные workflow-системы играют важную роль в обеспечении этого подхода.

Находясь на высшем уровне промежуточного программного обеспечения, workflow-системы позволяют ученым моделировать, выполнять, реконфигурировать и перезапускать свои составные программы числовых экспериментов.

Существует множество научных workflow-систем. Такие специализированные научные workflow-системы, такие как Discovery Net, Taverna workbench и Kepler, предоставляют пользователю среду визуального программирования, позволяющую конструировать потоки работ в виде графа. Каждое направленное ребро такого графа демонстрирует соединение выводящего потока одного приложения, входящего в поток, со входным потоком последующего.

Приведем список прочих известных workflow-систем: Bioclipse -- графическая рабочая среда с поддержкой скриптов для выполнения сложных вычислений; Discovery Net -- одна из самых первых научных workflow-систем; Ergatis -- интерфейс для создания и мониторинга рабочих потоков; Galaxy -- система, ориентированная на использование в геномике; OnlineHPC -- достаточно распространенный онлайн workflow-дизайнер и набор высокопроизводительных вычислительных инструментов; OpenMOLE -- научная workflow-система с интуитивно понятным набором настроек, начиная от многопоточных вычислений и заканчивая работой в среде Грид; Orange -- среда для анализа и визуализации данных с открытым кодом; Taverna -- научная workflow-система, работающая в облачной среде и объединяющая в себе основные качества таких систем, как Taverna и Galaxy.

По результатам проведенного анализа существующих средств особо перспективными следует считать распределенные системы на основе веб-сервисов с привлечением облачных ресурсов или использующие Грид-компьютинг для снятия ограничения на используемые вычислительные ресурсы.

Полноценное использование вышеупомянутых систем предполагает наличие детальных и глубоких знаний семантики каждого workflow-языка, в том числе понимания выполнения потоков данных в узлах и дугах workflow-графа, понимание функциональных эквивалентностей между workflow-паттернами, подобности типов данных и т.д. С практической точки зрения, требуется создание интегрирующих инструментов пользовательского уровня с интуитивно понятным интерфейсом пользователя, способных взаимодействовать с различными workflow-системами и их компонентами. Это предполагает разработку прототипа универсального веб-редактора потоков задач, который является ключевым элементом виртуального рабочего пространства пользователя, и проверка его совместной работы с набором существующих движков выполнения потоков задач (BPEL, Taverna и др.) на примере учебных заданий. Это позволит продвинуться в решении проблемы функциональной совместимости различных workflow-систем.

Література. 1. Barker A., Van Hemert J. Scientific Workflow: A Survey and Research Directions // 7th International Conference Parallel Processing and Applied Mathematics, PPAM 2007, Revised Selected Papers, Lecture Notes in Computer Science 4967. Gdansk, Poland: Springer Berlin / Heidelberg. – 2008. – pp. 746–753