Смирнов Кирилл Константинович,
Чернышев Георгий Алексеевич

УДК 004.65

УЧАСТИЕ В СОРЕВНОВАНИИ ACM SIGMOD КАК ВОЗМОЖНОСТЬ ДЛЯ СТУДЕНТОВ УГЛУБЛЕННО ИЗУЧИТЬ НЕКОТОРЫЕ АСПЕКТЫ БАЗ ДАННЫХ И ПРОГРАММНОЙ ИНЖЕНЕРИИ[1]

Аннотация

В настоящее время у студентов имеются богатые возможности для углубленного изучения различных областей ИТ: факультативы, летние школы, соревнования по программированию. В данной статье делается краткий обзор этих возможностей, а также описывается опыт участия в ежегодном соревновании по программированию наукоемких систем, проводимом при конференции ACM SIGMOD[2].

Ключевые слова: базы данных, acm sigmod contest, дополнительное образование.

1. ВВЕДЕНИЕ

Вузы, предоставляя возможность людям получить высшее образование, обеспечивают широкий выбор дисциплин для изучения. Обучение в вузе подразумевает фундаментальную подготовку по профильным предметам (в настоящей работе мы ограничимся такими дисциплинами как информатика и программная инженерия). Всё же по целому ряду причин [1] высокомотивированные студенты прибегают к различным дополнительным занятиям с целью более глубокого и детального освоения материала.

Существует множество мероприятий, с помощью которых студенты могут углубленно изучить интересующий их предмет, либо ознакомиться с чем-то новым. Перечислим некоторые из них.

1. Вневузовское обучение. В качестве примера можно привести обучение в Computer Science Center[3]. Это учебное заведение предлагает трехлетнюю программу обучения по трем направлениям. В целом обучение похоже на обучение на вечернем отделении: занятия проходят по вечерам, есть сессия, определенные правила сдачи курсов и т. д. Однако необходимо заметить, что правом выдачи дипломов государственного образца данное заведение не обладает.

2. Внутривузовские факультативы. Сюда можно отнести различные курсы, читаемые в университете, но не влияющие на академическую успеваемость студентов. Начиная с третьего курса, иногда даже со второго, заинтересованные студенты начинают посе-

[1] Работа выполнена при частичной поддержке РФФИ (грант 12-07-31050).

[2] http://www.sigmod.org/ .

[3] http://compscicenter.ru/ .

щать интересующие их спецкурсы и спецсеминары для участия в их работе.

3. Дистанционное обучение: курсы Coursera[1] и другие подобные. Такие курсы проводятся онлайн, включают в себя лекции и упражнения. Курсы Coursera насчитывают в данный момент более 200 различных курсов. Данный вид дополнительного образования может быть особенно полезен в случае, когда студент интересуется тематикой, а соответствующие курсы не читаются в университете.

4. Практические летние школы, такие как школа, проводимая кафедрой системного программирования математико-механического факультета СПбГУ совместно с ЗАО «Ланит-Терком» [2]. Условия занятий приближены к реальной работе в промышленной компании, что позволяет максимально повысить эффективность обучения. В рамках занятий студенты осваивают многие практики промышленного программирования, такие как обзор кода (code review), тестирование и прочие. Определенный упор делается на работу в команде, что сложно бывает организовать в рамках академических курсов, которые предполагают индивидуальную работу.

5. Теоретические летние школы как российские, так и международные. Среди первых можно отметить серию школ RuSSIR (школа, посвященная информационному поиску), школа Microsoft (тематика меняется ежегодно, варьируется от теоретической информатики до параллельного программирования) и многие другие. Международные школы отличаются тем, что там часто требуется оплата участия и проживания, хотя можно получить частичное или полное возмещение расходов. Кроме того, рабочий язык школы всегда английский (необходимо заметить, что вышеупомянутые российские школы также придерживаются этого правила). Примерами таких школ могут быть PROMISE и ESSIR (информационный поиск, базы данных) и другие.

6. Стажировки в компаниях, школы при компаниях. Последние отличаются от практических летних школ тем, что нацелены на подготовку и отбор кадров для последующего трудоустройства. Отличаются повышенными требованиями к участникам.

7. Различные соревнования по программированию, такие как ACM ICPC, Top Coder, ACM SIGMOD Programming Contest, Интернет-Математика и многие другие.

8. Прочие мероприятия, такие как Google Summer of Code. Цель данного мероприятия – реализация какого-либо программного проекта с открытым исходным кодом (open-source) или улучшение существующего. В отличие от соревнований, в данном случае студент может самостоятельно выбрать тему и задачу, которыми он хочет заниматься. Сюда же можно отнести и обычное участие в open-source проектах.

2. О СОРЕВНОВАНИЯХ ПО ПРОГРАММИРОВАНИЮ

По сравнению с другими видами дополнительного образования, соревнования по программированию обладают определенными плюсами:

– обладают продуманной, четко определенной постановкой задачи;

– практический подход может быть интересен студентам больше, нежели теория;

– участие в соревновании дает студентам дополнительную мотивацию.

Соревнования по программированию можно разделить на две группы: общего алгоритмического характера и определенной тематической направленности. К первой группе можно отнести ICPC и Top Coder. Рассмотрим их отличия от тематических на примере сравнения с ICPC (см. табл. 1). Конечно, некоторые детали у разных соревнований могут отличаться (например, в соревнованиях Top Coder участие индивидуально), это не жесткая схема. Однако можно выделить основной набор отличий.

Обсудим наиболее важные моменты:

1. Характер задач определяется длительностью соревнования. Так, в условиях ICPC невозможно решение задач, требующих

[1] https://www.coursera.org/ .

Табл. 1. Отличия соревнований общего характера от тематических, на примере сравнения ACM ICPC и ACM SIGMOD Contest

	Соревнование ACM ICPC	**Соревнование ACM SIGMOD**
Длительность	Несколько часов	Несколько месяцев
Предметная ориентация	Классические алгоритмы области computer science	Специализированные алгоритмы, набор тем по базам данных
Количество задач	Несколько	Одна большая
Приемка	Закрытое автоматическое тестирование	Закрытое тестирование, в сложных случаях возможны подсказки
Порядок приемки	Набор тестов для проверки корректности	Набор тестов для проверки корректности, измерение производительности под нагрузками
Количество команд, участвующих от университета	Ограничено	Не ограничено
Набор тестов	Не известен	Не известен
Разрешенные инструменты	IDE, встроенные отладчики	Более широкий набор отладчиков и профайлеров, система контроля версий, система управления проектом, практически нет ограничений.
Просматривается ли код участников	В исключительных случаях	Просматривается всегда, проверяется на предмет нарушения правил
Ранжирование команд	Количество сданных задач, штрафное время	Какая-либо метрика производительности (пропускная способность системы, время отклика, время обработки набора запросов и т. д.)

больше, чем несколько десятков человеко-часов труда. Отсюда проистекают ограничения на структурную сложность решаемой задачи. Задачи носят «рафинированный» характер, подавляющее большинство из них уже имеет известное алгоритмическое решение.

2. Вследствие предыдущего пункта, накладываются ограничения и на используемые средства программной инженерии. Для успешной сдачи задачи на ICPC требуется обеспечить работоспособность относительно небольшой программы на тщательно продуманном наборе тестов. Для этого не требуется ни владение серьезным отладочным инструментарием (за исключением пошаговой отладки в IDE), ни системой контроля версий. Более того, это запрещено правилами.

3. Еще одно следствие длительности соревнования и наличия одной большой задачи – принципиально иное распределение обязанностей в команде.

4. ACM SIGMOD Contest ориентируется на специфическую предметную область (базы данных), в то время как соревнование ICPC ориентируется на более общие фундаментальные основы информатики.

5. Разный подход к тестированию, существование правильного решения. В соревно-

вании ACM SIGMOD Contest не существует единственно правильного, задуманного авторами решения, так как задача представляет собой научную проблему. В финале практически всегда присутствуют несколько различных классов подходов. Из этого проистекает существенная разница в подходах к тестированию.

Таким образом, участники соревнований ICPC овладевают обширными познаниями в области теоретической информатики, изучают на практике классические алгоритмы и структуры данных. Люди, обладающие такой подготовкой, представляют особую ценность на мировом рынке производства наукоемкого ПО. Однако такая широта охвата влечет за собой недостаточную глубину: сложные и узкоспециализированные структуры данных (например GiST и R-дерево) будут обойдены вниманием, так как их просто не успеют написать и отладить за те несколько часов, что идет соревнование. В целом, как следует из вышеприведенных пунктов, нишевые соревнования более приближены к задачам реального мира.

Почему имеет смысл обратить внимание на тематические, в частности, на ACM SIGMOD Contest:

– соревнование предлагает наукоемкие и актуальные задачи;

– соревнование предлагает сложные с технической точки зрения задачи;

– знакомство с научным сообществом, возможность участия в одной из лучших в мире научных конференций.

В результате участия в этом соревновании студенты получат следующие навыки:

1. Изучат на практике, как реализуется один из компонентов СУБД. Современный подход к проведению практических занятий по базам данных во многом ориентирован на подготовку прикладных программистов. Это можно объяснить многими причинами, например нехваткой часов или техническими трудностями. Участие в данном соревновании может помочь перешагнуть пропасть между прикладным программированием и системным, привнесет понимание того, как работает тот или иной компонент досконально.

2. Студенты обучатся избранным аспектам системного программирования. В данном соревновании очень сильны данные аспекты, их, по разным причинам, трудно в совершенстве отработать на практике в ходе обучения в университете. К ним можно отнести многопоточное программирование, использовавшееся во всех прошедших соревнованиях, или работу с сетью.

3. Студенты могут приобрести опыт командной разработки больших проектов, отработать важные навыки в процессе разработки программного обеспечения: планирование задач, этику общения по электронной почте, использование системы контроля версий, модульного тестирования и отладки. Эти аспекты также трудно преподать на практике в ходе университетских курсов.

4. Студенты обучатся работе с научной литературой. В ходе соревнования студентам будет необходимо выбирать подходы к решению, знакомиться с актуальными наработками путем чтения литературы по тематике соревнования. Навык такой работы преподается в университетах, однако чем раньше его приобретут интересующиеся студенты, тем лучше. Немаловажную роль играет и его закрепление на практике, в «боевых условиях».

5. Кроме того, важно отметить возможность приобрести навыки делового общения и переписки на иностранном языке. Участники и организаторы соревнования – члены ведущих мировых исследовательских групп по базам данных и информационному поиску. По ходу соревнования происходят оживленные обсуждения, участниками задаются вопросы, организаторы отвечают и поясняют условие или детали задачи.

3. СОРЕВНОВАНИЕ ACM SIGMOD CONTEST

3.1. О СОРЕВНОВАНИИ

ACM SIGMOD (ACM Special Interest Group on Management of Data) – сообщество, занимающееся принципами, методами и приложениями СУБД и технологией управления данными в целом. Оно было основано в 1976 году, с этим сообществом связаны

имена таких известных ученых, как Альфреда Ахо, Питера Чена, Джеффри Ульмана, Майкла Стоунбрейкера и многих других. Конференция SIGMOD – флагманская конференция данного сообщества, и при этом, это одна из наиболее престижных научных конференций по базам данных в мире. При этой конференции и проводится обсуждаемое соревнование.

ACM SIGMOD Programming Contest – международное соревнование студентов и аспирантов вузов, посвященное тематике баз данных. Впервые данное соревнование было проведено в 2009 году и с тех пор проводится ежегодно. Организаторами выступает каждый год новая группа ученых и преподавателей, при этом существует постоянный программный комитет.

Тематика данного соревнования покрывает довольно широкий спектр тем, относящихся к базам данных. К ним относятся: различные методы индексирования, модели исполнения запросов, обеспечение обработки транзакций и другое.

Каждый год перед участниками ставится задача реализовать какой-либо компонент или подсистему СУБД. Рассматриваемая задача – это всегда научная проблема, при этом обладающая определенными техническими трудностями при реализации.

Помимо выявления лучших студентов и аспирантов, целью данного соревнования ставится создание полнофункциональной СУБД. Складывая вместе результаты каждого года, планируется шаг за шагом строить систему (реализация-победитель каждого года публикуется на сайте под лицензией MIT). Конечно, данная система никогда не сможет сравниться по производительности или другим характеристикам с коммерческой СУБД или даже с СУБД с открытым кодом наподобие PostgreSQL или MySQL. Однако она сможет соперничать с ними на поле систем для обучения студентов. Вышеупомянутые системы обладают определенной архитектурной сложностью и оптимизированы для выполнения своих задач. Основной массе студентов этот код недоступен для понимания в сжатые сроки обучения.

Финалисты получают возможность поучаствовать в работе конференции и прослушать доклады. Самое главное – они получают возможность выступить со стендовым докладом (poster). Кроме того, участники получают возмещение затрат.

В соревновании заинтересованы крупные зарубежные IT компании (Microsoft, Amazon, SAP), оно проводится при их поддержке.

3.2. РАСПОРЯДОК СОРЕВНОВАНИЯ

Это ежегодное соревнование, которое начинается поздней осенью или зимой, заканчивается обычно весной-летом. Конкретные даты сильно зависят от дат конференции, так как привязаны к ним, а даты конференции плавающие. Каждый год соревнование проходит несколько фаз, попытаемся их выделить и приблизительно датировать (они также зависят от дат конференции).

1. Объявление. Происходит, как уже было отмечено, поздней осенью или зимой. На указанный момент становится известна тема в широком смысле, направление.

2. Объявление точной постановки задачи – конец декабря-февраль. Постановка задачи объявляется полностью, представляется интерфейс, который необходимо реализовать.

3. Появление тестового стенда. В процесс проведения соревнования вводится специальное оборудование с известной конфигурацией, на котором будут испытываться решения. Участники могут отправлять свои решения на проверку и видеть их оценку. Начинает вестись открытый рейтинг команд.

4. Окончание первого тура. Обычно происходит за месяц-два до начала конференции. Участники обязаны предоставить исходные коды своих реализаций для оценки и выбора финалистов. С этого момента регистрация новых участников запрещена, рейтинг команд закрывается.

5. Окончание второго тура. Происходит за несколько дней или недель до начала конференции. После этого подача реализаций невозможна.

6. Объявление победителей и награждение происходит на самой конференции.

3.3. О ПОДГОТОВКЕ К СОРЕВНОВАНИЮ

Авторы данной работы готовили команду студентов третьего курса к соревнованию 2012 года. Подготовка велась с разной степенью интенсивности, в целом охватывался период чуть меньше года. Для подготовки к соревнованию нами было сделано следующее.

1. На лето выдан к прочтению материал из учебников по базам данных. Дело в том, что курс баз данных находится в программе старших курсов, и к моменту начала соревнования он еще не был прочитан.

2. На лето выдано несколько статей по базам данных на английском языке, с целью как ознакомления с современным материалом, так и с целью получения практики чтения англоязычной литературы.

3. Прочитан дополнительный курс лекций по базам данных. Основной упор делался именно на практических аспектах реализации СУБД и подобных систем, материал подбирался из современных статей или практических глав учебников.

4. Прочитан дополнительный курс лекций по программной инженерии: ведение проектов, углубленное изучение систем контроля версий, тестирования и отладки. Рассмотрены различные средства отладки и профилирования из семейства инструментов динамического анализа приложений Valgrind, объяснено, как пользоваться простыми инструментами.

5. Прочитаны дополнительные лекции, посвященные системному программированию для платформы Linux: многопоточное программирование, работа с сетью, файловый ввод-вывод и др.

3.4. ПРОБЛЕМЫ, С КОТОРЫМИ ВОЗМОЖНО СТОЛКНУТЬСЯ

Перечислим ряд трудностей, с которыми мы столкнулись и как участники, и как тренеры команды.

1. К сожалению, отсутствует необходимая литература на русском языке. Безусловно, в переводе существуют отличные учебники, такие как К. Дж. Дейт «Введение в системы баз данных» или серия книг Ульмана, однако тут возникают проблемы. Первая проблема – это довольно старые учебники, выдержавшие множество переизданий. Они делают основной упор на классических темах. В соревновании такие темы редко встречаются, знаний, предлагаемых этими учебниками, не хватает. Вторая проблема – эти учебники широкого профиля, то есть там разбираются сразу несколько слабосвязанных тем. Тематические учебники практически полностью отсутствуют в переводе на русский язык. Так, например, касательно области обработки транзакций, ни книга Герхарда Вейкума «Transactional Information Systems: Theory, Algorithms, and the Practice of Concurrency Control and Recovery», ни книга Джима Грея «Transaction Processing: Concepts and Techniques» (знаменитые тематические учебники) на русский язык не переведены. Третья проблема –слабо представленный практический компонент. Выход из данной ситуации – работа с англоязычными учебниками и, что самое главное, – со статьями.

2. Требуется хорошее аппаратное обеспечение. Это может быть как кластер рабочих станций, так и одна многопроцессорная машина. Или, например, новые, дорогостоящие компоненты. Так, в 2011 году требовался твердотельный жесткий диск (solid state drive, SSD). Наличие такого обеспечения желательно, но не обязательно – как было отмечено выше, организаторы предоставляют указанное оборудование. Однако на тестовый стенд образуются очереди, или он может выйти из строя на длительное время, как, например, было в 2010 году (по вине именно нашей команды).

3. Серьезные требования к практической подготовке студентов. Как решать эту проблему, нами было описано выше.

3.5. ОБ ИТОГАХ СОРЕВНОВАНИЯ

Команда, состоявшая из авторов настоящей статьи, бывших на тот момент аспирантами последнего курса математико-механического факультета СПбГУ, в соревновании 2010 года заняла третье место среди трех десятков команд. Темой была задача пост-

роения исполнителя и оптимизатора запросов для распределенной СУБД. Детали нашего решения (описание прототипа) и некоторая конкретика соревнования представлена в [3].

Команда студентов-третьекурсников, которую готовили авторы данной статьи, в соревновании 2012 году заняла пятое место на публичных тестах. В качестве темы было заявлено построение многомерного индекса, детали постановки задачи и нашего решения представлены в [4-6]. Необходимо сказать, что среди участников преобладают аспиранты, а не студенты. Например, в 2010 году среди пяти команд-финалистов была только одна команда студентов. Кроме того, были выявлены серьезные проблемы в организации соревнования.

1. Плохо подготовленные тесты, проверявшие корректность реализации. Не был проверен тривиальный случай, когда данные не добавлялись в структуру данных. Такое может происходить, например, при ошибках в реализации. Этот случай был найден нашей командой, о нем было сообщено за несколько дней до конца подачи, организаторы добавили тест и произвели повторное тестирование всех решений.

2. Сорван график соревнования, организаторы были вынуждены сделать продление в последние дни, перед концом подачи. Это – исключительная ситуация, никогда ранее не происходившая. К сожалению, она сыграла против нас, так как на это время у студентов приходилась важная аттестация в вузе.

3. Из-за сорванного графика соревнования и непродуманных правил организаторы были вынуждены пойти на беспрецедентный шаг – производить исправление реализаций, в случае, если они не заработают на их тестах.

В целом можно сказать, что организаторы соревнования этого года справились с задачей подготовки и проведения соревнования хуже всех предыдущих.

Таким образом, итог соревнования 2012 года должен рассматриваться как безусловный успех.

4. ЗАКЛЮЧЕНИЕ

В данной работе мы перечислили несколько возможностей получения дополнительного образования студентами. Была рассмотрена возможность участия в различных соревнованиях по программированию, и особенно подробно рассмотрено соревнование ACM SIGMOD Contest, произведено сравнение с другими соревнованиями. Были рассмотрены распорядок соревнования и правила, описаны мероприятия по подготовке студентов, представлены возможные трудности.

В заключение можно сказать, что участие в соревновании дало следующие результаты:

– проведены актуальные исследования, по результатам которых опубликованы статьи в реферируемых журналах и на конференциях, еще несколько готовятся к публикации;

– студенты приобрели и закрепили на практике навыки коллаборативной разработки проектов;

– студенты изучили отдельные аспекты системного программирования;

– студенты приобрели навыки работы с научной литературой на иностранном языке, приобрели навыки делового общения и переписки на иностранном языке;

– студенты на практике изучили, как реализуется один из компонентов СУБД, какие могут возникать технические и научные проблемы при реализации.

Литература

1. *Терехов А.Н.* Что такое программная инженерия // Программная инженерия, 2010. № 1. С. 40–45.

2. *Брыксин Т.А.* Студенческие проекты по программированию как средство формирования профессиональных навыков // Системное программирование. Том 6, вып. 1: Сб. статей / Под ред. А.Н. Терехова, Д.Ю. Булычева. СПб.: Изд-во СПбГУ, 2011. С. 116–135.

3. *Смирнов К.К., Чернышев Г.А.* Сетевые и многопоточные аспекты архитектуры распределенных СУБД // Программные продукты и системы, 2011. № 1 (март). С. 164–169.

4. *Ерохин Г.А., Чернышев Г.А.* Экспериментальное сравнение алгоритмов разделения вершин в R-дереве на различных данных / Материалы третьей межвузовской научной конференции по проблемам информатики СПИСОК-2012, 2012.

5. *Федотовский П.В., Чернышев Г.А., Смирнов К.К.* Реализация уровня изоляции Read Committed для древовидных структур данных / Материалы третьей межвузовской научной конференции по проблемам информатики СПИСОК-2012, 2012.

6. *Чередник К.Е., Смирнов К.К.* Динамическое распределение памяти в многопоточном обработчике транзакций / Материалы третьей межвузовской научной конференции по проблемам информатики СПИСОК-2012, 2012.

Abstract

Nowadays, students have a lot of opportunities for deep study of various disciplines of IT: extracurricular courses, summer schools, programming contests. In this paper we make a short survey of such opportunities and describe details of our participation in the annual ACM SIGMOD Programming contest, held by the SIGMOD conference and devoted to development of science-intensive database applications.

Keywords: database systems, acm sigmod contest, extracurricular education.

Смирнов Кирилл Константинович,
программист СПбГУ,
kirill.k.smirnov@math.spbu.ru,

Чернышев Георгий Алексеевич
ассистент математико-механического факультета СПбГУ,
chernishev@gmail.com .