Институт прикладной математики им. М.В.Келдыша РАН

МЕХАНИЗМЫ УПРАВЛЕНИЯ РАЗДЕЛЯЕМЫМИ КОМПЬЮТЕРАМИ В ГРИДЕ

П.С. Березовский, В.Н. Емельянов,
В.Н. Коваленко, Э.С. Луховицкая

Сложности управления неотчуждаемыми ресурсами

Неотчуждаемые ресурсы должны использоваться в разделяемом режиме, при этом возникают следующие проблемы:

- необходим механизм разделения ресурсов с соблюдением автономности прав владельца
- усложняется задача планирования, так как необходимо обеспечить выполнение задания в срок на недетерминированных ресурсах

Работающие инфраструктуры

Наиболее широкое распространение получила система BOINC и реализованные на её основе проекты – серия @Home. В этих проектах задействовано около 1 млн. компьютеров (0,1% от общего числа ПК в мире).

Самым первым и наиболее популярным проектом является SETI@Home. В рамках этого проекта*:

- подключено – 330 000 компьютеров (90% из них под Windows)
- среднее время непрерывного участия – 91 день
- общая производительность – 149,8 TFlops

Суммарная производительность по всем проектам – 1084,16 TFlops

* David P. Anderson and Gilles Fedak. The Computational and Storage Potential of Volunteer Computing. IEEE/ACM International Symposium on Cluster Computing and the Grid, Singapore, May 16-19, 2006

Потенциал некластеризованных компьютеров

- Сколько времени компьютер занят, а сколько простаивает?

- Исследование* показывает, что 46% рабочего времени компьютеры находятся в состоянии занятости, но на 76% таких периодов загрузка процессора не превышает 10%, то есть большинство периодов занятости относится к работе с клавиатурой.

По результатам исследования загрузки процессора:

- низкая загрузка (меньше 10%) составляет 82% времени
- средняя загрузка (от 10 до 80%) – 13%
- высокая загрузка (более 80%) – 7%

Вывод: процессоры рабочих станций теряют значительную долю мощности более 90% времени.

* Kyung Dong Ryu and Jeffrey K. Hollingsworth. Exploiting Fine-Grained Idle Periods in Networks of Workstations. IEEE Transactions On Parallel And Distributed Systems, Vol. 11, No. 7, July 2000.

Влияние внешних задач на работу пользователя*

	Процессор	Память	Диск
Word	4.35	*	4.20
PowerPoint	1.17	0.64	4.65
IE	1.20	0.55	3.11
Quake	0.64	0.55	1.19
Всего	1.47	0.58	2.97

Критерий оценки – число процессов, одновременно обращающихся к ресурсу.

Оценка чувствительности приложений к разделению ресурсов.

	Процессор	Память	Диск	Всего
Word	Низ	Низ	Низ	Низ
PowerPoint	Ср	Низ	Низ	Ср
IE	Ср	Ср	Выс	Ср
Quake	Выс	Ср	Ср	Выс
Всего	Ср	Низ	Низ	

* Gupta, A., Lin, B., and Dinda, P. A. Measuring and understanding user comfort with resource borrowing. In Proceedings of the 13th IEEE International Symposium on High Performance Distributed Computing (HPDC 2004), June 2004

Механизмы разделения ресурсов

- Внешнее задание выполняется только в случае простоя компьютера (BOINC, Condor)
  - ✓ локальные задания получают необходимые ресурсы в полном объёме
  - ✓ владелец может определять условия запуска внешних заданий
    - время запуска
    - объём доступных ресурсов
    - …
- Заимствование циклов – Fine-Grained Cycle Steeling
  - ✓ внешнее задание и локальные выполняются параллельно
  - ✓ разработанные методы позволяют занимать свободные ресурсы, не оказывая влияния на работу владельца
    - метод Linger-Longer* способен продуктивно использовать свыше 90% свободных циклов процессора (внешние задания запускаются с минимальным приоритетом и не мешают выполнению локальных).

* Kyung Dong Ryu, Jeffrey K. Hollingsworth. Exploiting Fine-Grained Idle Periods in Networks of Workstations. IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 11, NO. 7, JULY 2000

Методы планирования

Недетерминированность ресурсов не позволяет гарантировать выполнение задания в срок.

Предлагаются два подхода:

- использование разделяемых компьютеров совместно с выделенными, которые гарантируют выполнение задания в срок
- планирование на основе статистических характеристик разделяемых компьютеров, что позволяет повысить качество планирования и уменьшить число отказов

Планирование на основе статистических характеристик

- Известные модели ориентированы на предсказание поведения в пределах коротких отрезков времени (1-15 сек), что подходит только для краткосрочного управления разделением компьютера

- Мы ставим задачу исследовать, что может дать учёт статистических характеристик для улучшения планирования заданий с продолжительным временем обработки

Статистические данные для планирования

- Номинальная производительность компьютера (G)
- Эффективная производительность компьютера – доля процессорного времени, предоставляемая гриду (H)

$$h=(T_1/T)*G$$

  - ✓ T_1 – количество времени процессора, полученного внешним заданием
  - ✓ T_1 – не используемое время процессора, если нет внешнего задания

- Средняя длина времени во включенном состоянии (L_A)
- Средняя длина времени в выключенном состоянии (L_{NA})
- Время выполнения задания на эталонном компьютере (W)
- Предельный срок завершения задания (D)

Оценка времени выполнения задания

t_0 – текущее время

t_{on} – время последнего включения компьютера

Если компьютер не выключается за время t, внешнее задание получит G*H*t нормированного процессорного времени.

Тогда время выполнения задания вычисляется следующим образом:

$$T_{finish} = \begin{cases} \frac{W}{G*H}, & \text{если } T < t_{on} + L_A - t_0 \\ \frac{W}{G*H} + L_{NA}, & \text{в противном случае} \end{cases}$$

Задача планирования

- Недетерминированность компьютеров ограничивает круг алгоритмов планирования простейшими (FCFS, List Scheduling, EDF), однако учёт статистических характеристик может позволить улучшить показатели планирования
- Задача планирования решается в следующих условиях:
  - ✓ выполняется распределение потока заданий $\{J_i, i=1,2,\ldots\}$ по компьютерам $\{C_j, j=1,M\}$
  - ✓ задание J_i определяется парой $\{W_i, D_i\}$ и получает ресурсы компьютера в разделяемом режиме
  - ✓ максимальное время занятия компьютера ограничено D_i
  - ✓ компьютеры характеризуются параметрами $\{G_j, H_j, L_{Aj}, L_{NAj}\}$

Алгоритм планирования

- Предполагается, что компьютеры, объединённые в пул, отличаются по своим статистическим характеристикам. Учёт этого обстоятельства позволяет уменьшить число отказов (заданий, превысивших срок выполнения)

- Рассматриваются два алгоритма:
  - ✓ FCFS – задание распределяется на любой свободный компьютер
  - ✓ модифицированный FCFS – задание распределяется только на тот компьютер, где оно может окончиться в срок

- Производится моделирование для синтетических данных, описывающих поток заданий и характеристики компьютеров. Для двух алгоритмов сравнивается число отказов при различных значениях отклонения предсказанной производительности компьютера от реальной, а также стабильность числа отказов в зависимости от числа компьютеров и заданий.

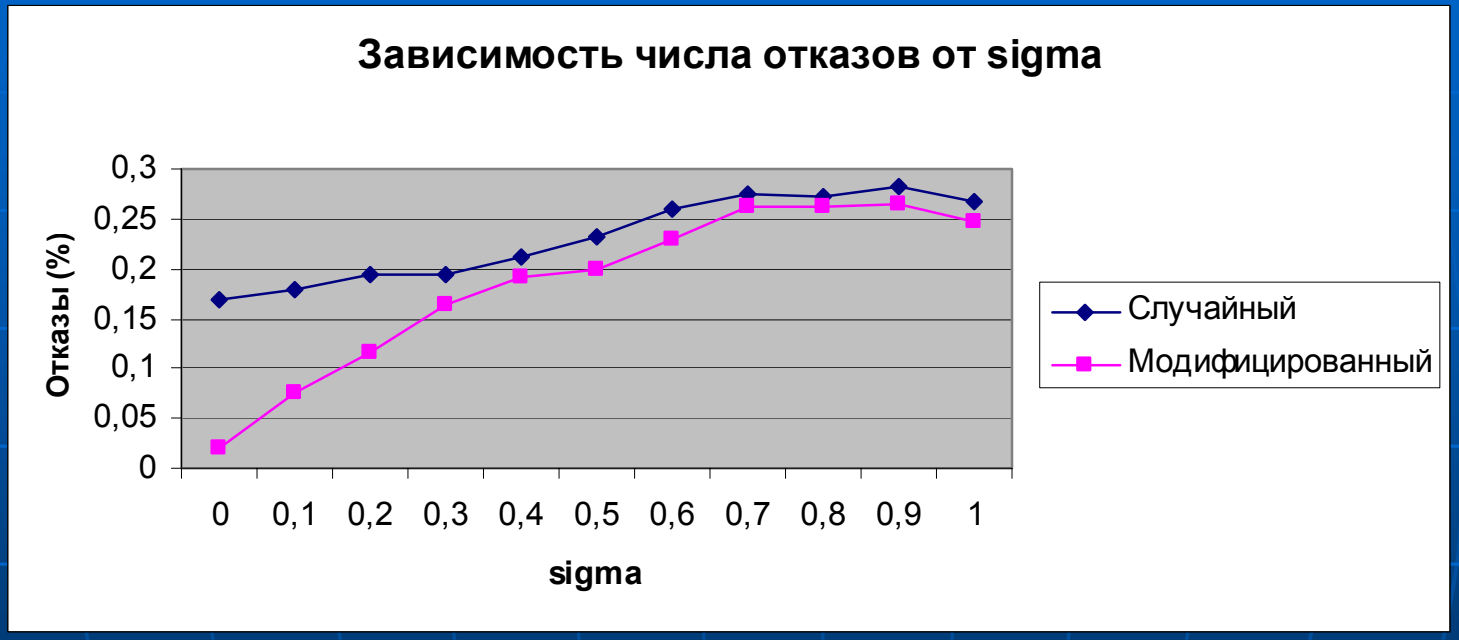
Моделирование

Условия моделирования*:

- Эффективная производительность компьютеров $H \in [H_{min}, H_{max}]$, $H_{min} = 0.2$, $H_{max} = 1$
- Длина задания $W \in [150, 750]$
- Время поступления заданий $ST \in [0, TS]$
- Предельный срок выполнения задания
  $D = ST + \alpha * W$
  $\alpha \in [\alpha_{min}, \alpha_{max}]$, $\alpha_{max} = 1/H_{min}$
- Количество компьютеров $NC \in [10, 100]$
- Количество заданий $NJ = m*NC$, $m \in [10, 100]$

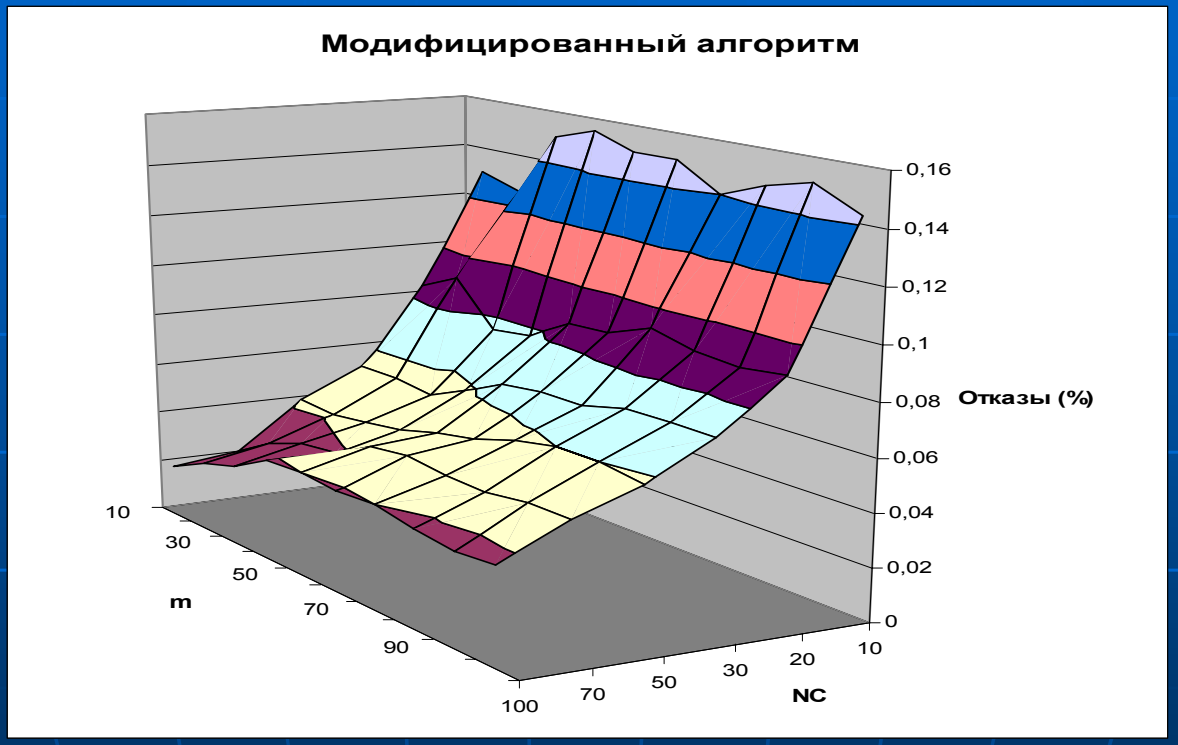
* Значения параметров равномерно распределены на соответствующих отрезках.

Результаты моделирования (1)



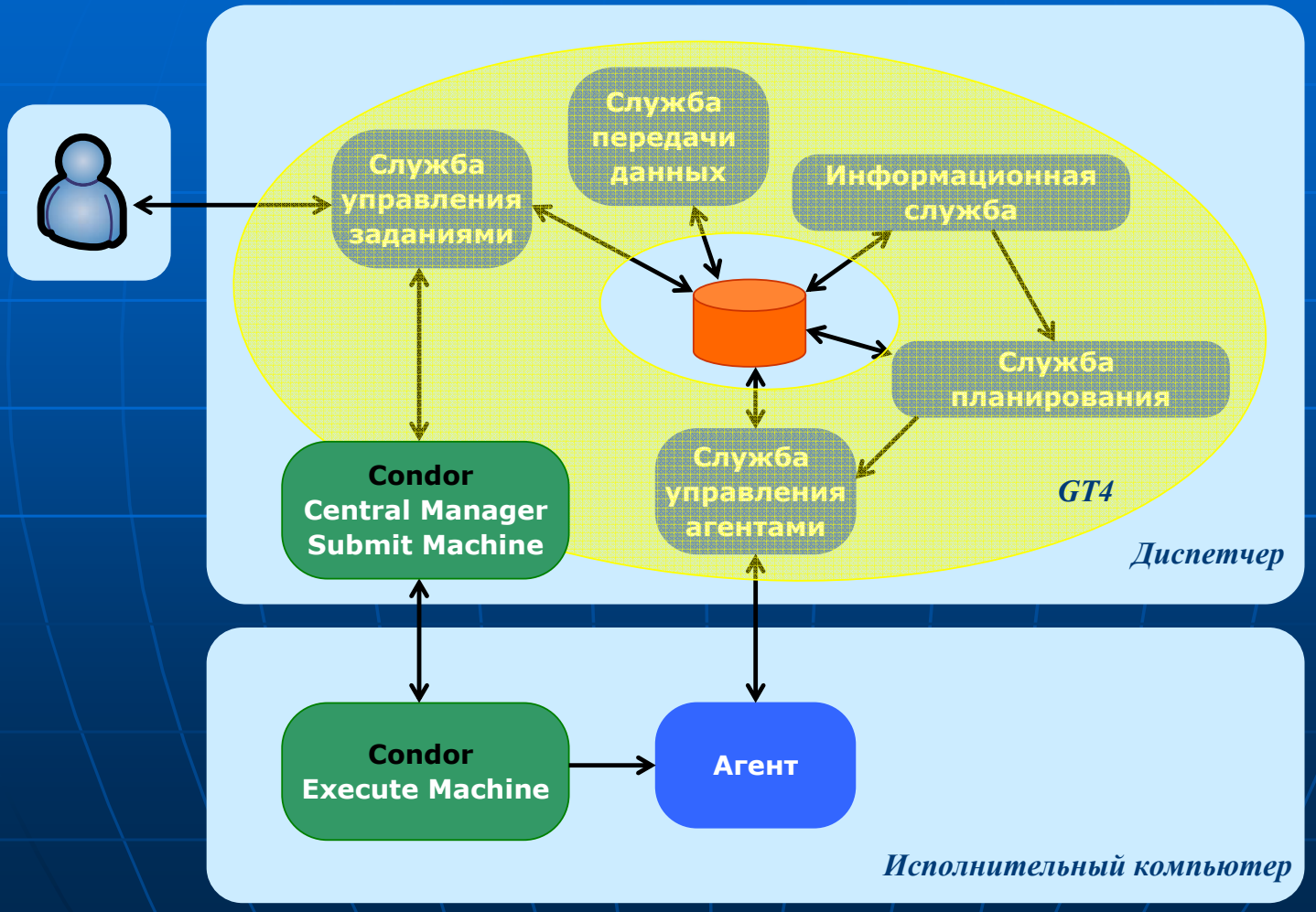
Результат: Наличие информации об эффективной производительности компьютера существенно уменьшает число отказов даже при 25% отклонении от реальной производительности.

Результаты моделирования (2)

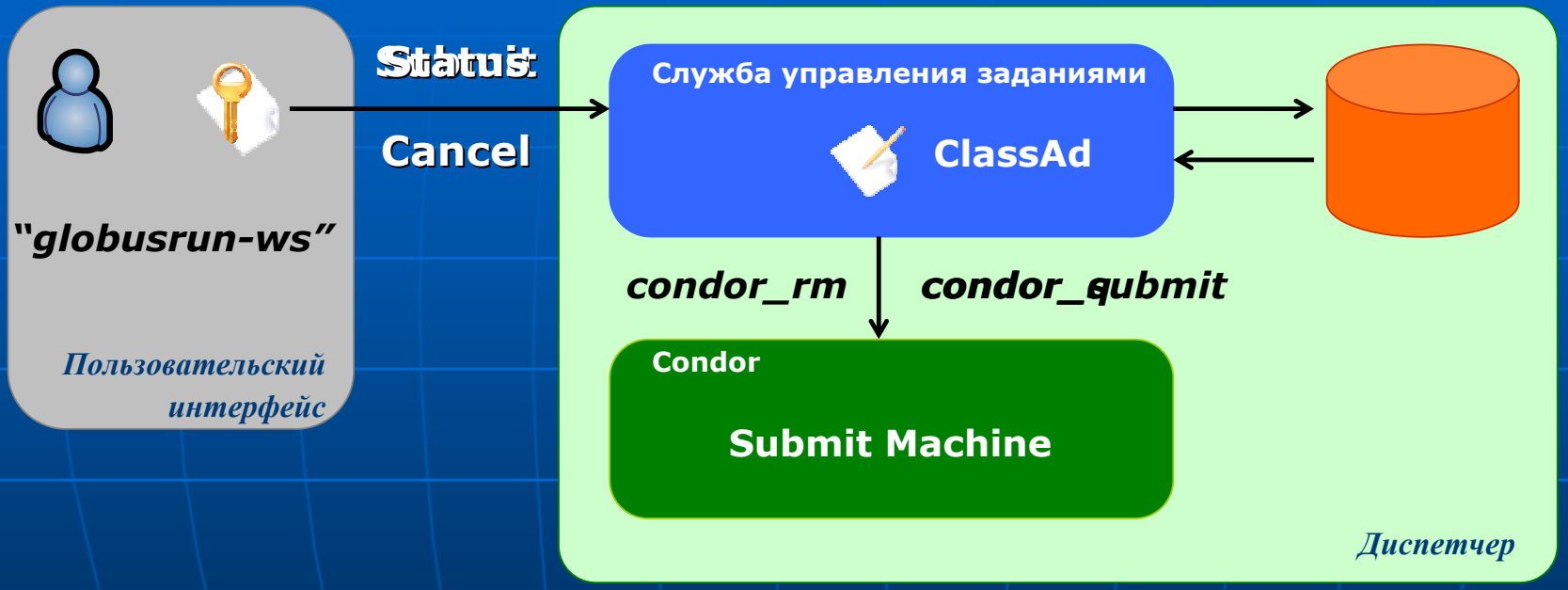


Результат: Модифицированный алгоритм адекватно реагирует на увеличение числа заданий и изменение количества исполнительных компьютеров.

Архитектура системы диспетчеризации



Служба управления заданиями и пользовательский интерфейс



- Расширенное описание задания (walltime и deadline)
- Формирование ClassAd и запуск на выбранном компьютере средствами Condor

Агент и служба управления агентами

- Функции агента:
  - ✓ Регистрация компьютера в системе и снятие с регистрации
  - ✓ Сбор информации о компьютере
  - ✓ Сбор информации о выполняющемся задании
- Служба получает от агента и реализует обработку следующих типов сообщений:
  - ✓ Регистрация компьютера
  - ✓ Снятие с регистрации
  - ✓ Подключение
  - ✓ Отключение
  - ✓ Получение периодического отчёта

Периодический отчёт агента

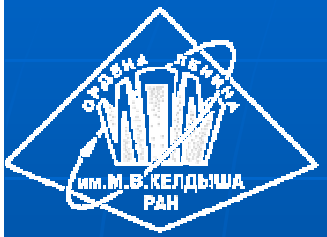
- Характеристики компьютера
  - ✓ Процессор (архитектура, производительность)
  - ✓ Операционная система
  - ✓ Оперативная память
  - ✓ Дисковое пространство
- Объём свободных ресурсов
  - ✓ Доля свободного процессорного времени за некоторый интервал времени
- Информация о запущенном задании
  - ✓ События, возникающие в процессе выполнения задания (запуск, завершение, аварийное завершение)
  - ✓ Доля процессорного времени, полученного заданием за определённый период времени

Недостатки системы Condor

- Узкое место системы – пользовательский интерфейс (машина запуска). Для каждого выполняющегося задания на машине запуска стартует отдельный процесс, который завершается только после окончания выполнения задания, что делает систему слабо масштабируемой.
- Планирование ограничено обеспечением запуска задания. Нет гарантии того, что задание будет выполнено в любой назначенный срок.

Спасибо за внимание!

Контакты



Институт прикладной математики им. М.В.Келдыша РАН

Россия, 125047, Москва, Миусская пл. 4; тел. (495) 250-79-82

- Березовский П.С bps@keldysh.ru
- Коваленко В.Н. kvn@keldysh.ru

Работы ИПМ в области грида доступны на сайте http://www.gridclub.ru

gridclub.ru