Архитектурные аспекты решения задачи фильтрации документов на потоке запросов

Федотовский Павел, 445 гр.

Научные руководители:
Чернышев Г.А.
Смирнов К.К.

Кафедра системного программирования, СПбГУ, 2013

Введение

- Нечеткий поиск – поиск всех строк, близких к данной с учетом заданной метрики
  - Поисковые системы
  - Биоинформатика

- Соревнования, Работы, Прототипы
  - SIGIR, SIGMOD...
- ACM SIGMOD Programming Contest 2013

Контекст работы

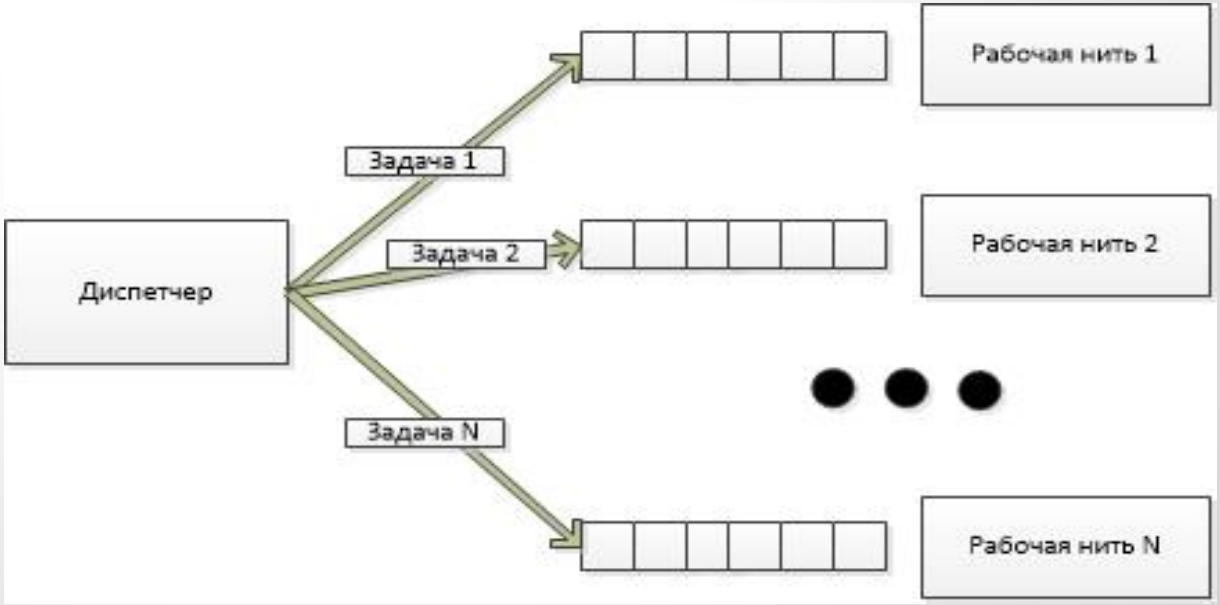
- Прототип высокопроизводительной многопоточной системы обработки документов
  - На вход поступают документы и запросы
  - Нужно найти все запросы, удовлетворяющие документу
  - Запрос подходит документу, если для каждого слова в запросе есть соответствие в документе
  - C++, GNU/Linux

Цель работы

- Обеспечить равномерную загрузку вычислительных узлов (24 ядра)
- Задачи
  - Изучить принципы организации нитей и реализовать подходящий
  - Выбрать разбиение крупных задач на подзадачи для нитей и реализовать

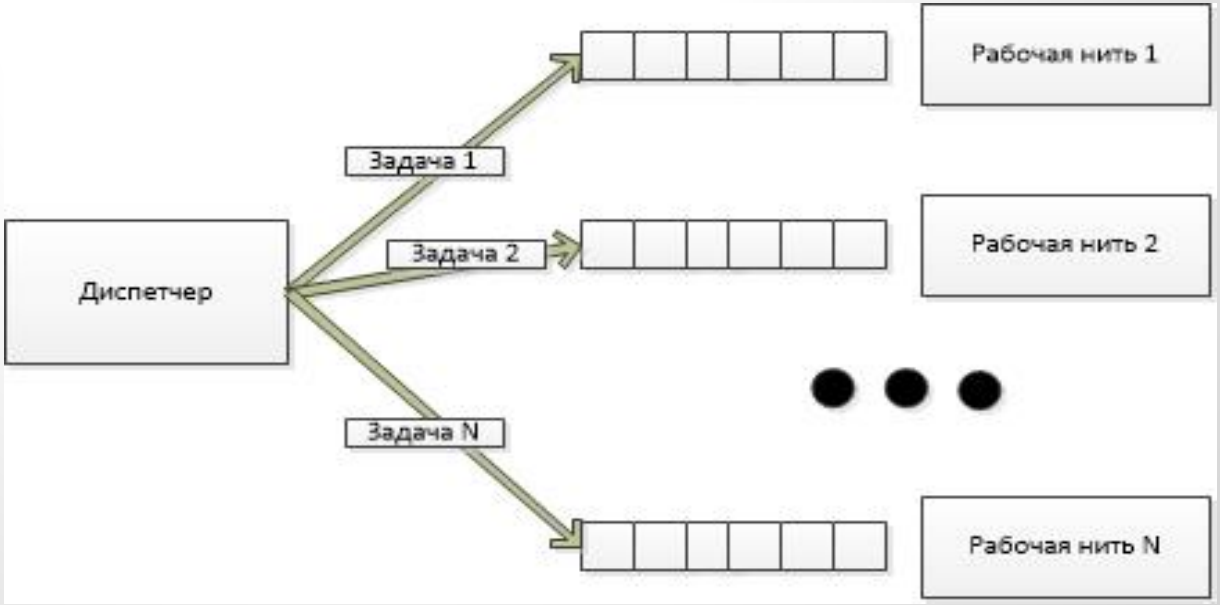
Архитектура

- **Round-Robin**
- Work Stealing
- Common Queue
- Thread Pool



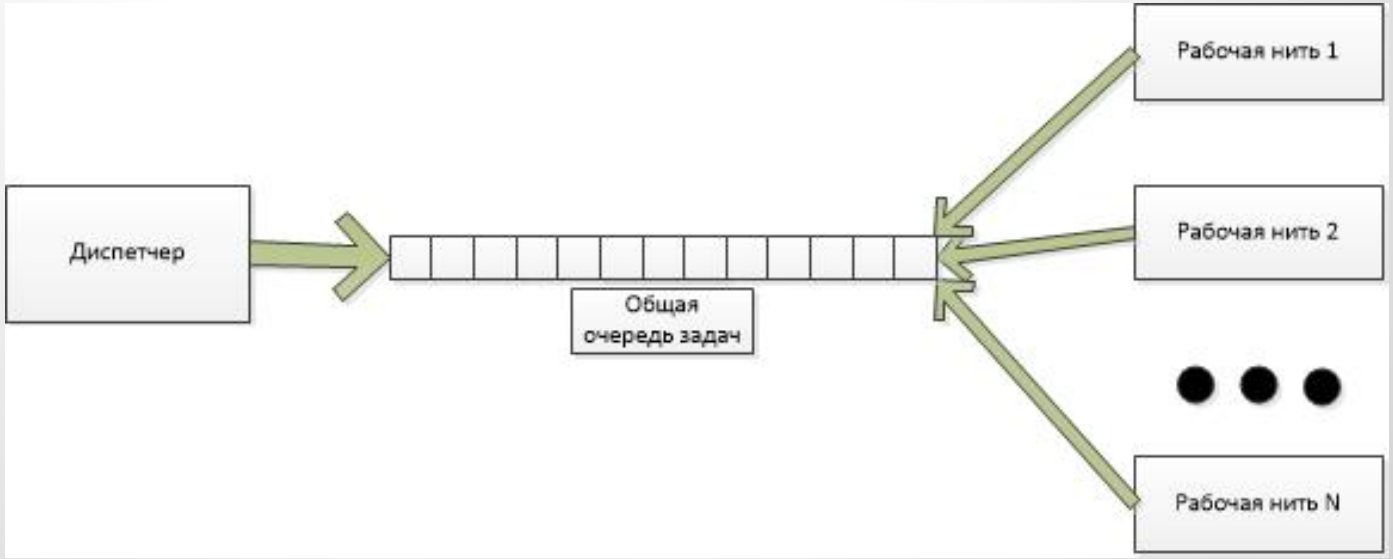
Архитектура

- Round-Robin
- **Work Stealing**
- Common Queue
- Thread Pool



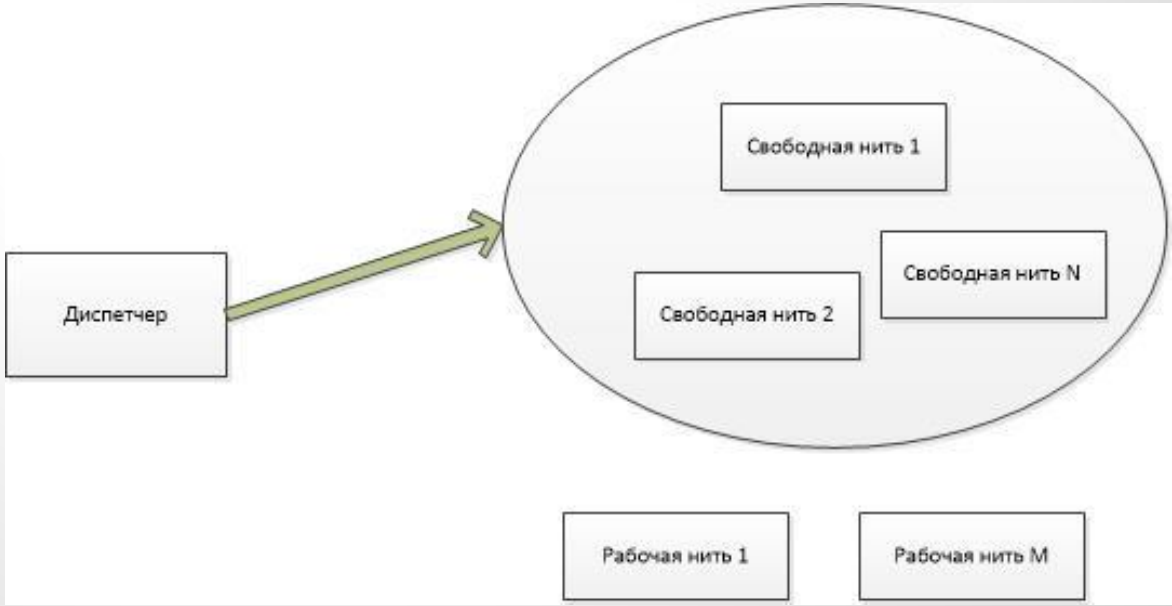
Архитектура

- Round-Robin
- Work Stealing
- **Common Queue**
- Thread Pool



Архитектура

- Round-Robin
- Work Stealing
- Common Queue
- **Thread Pool**

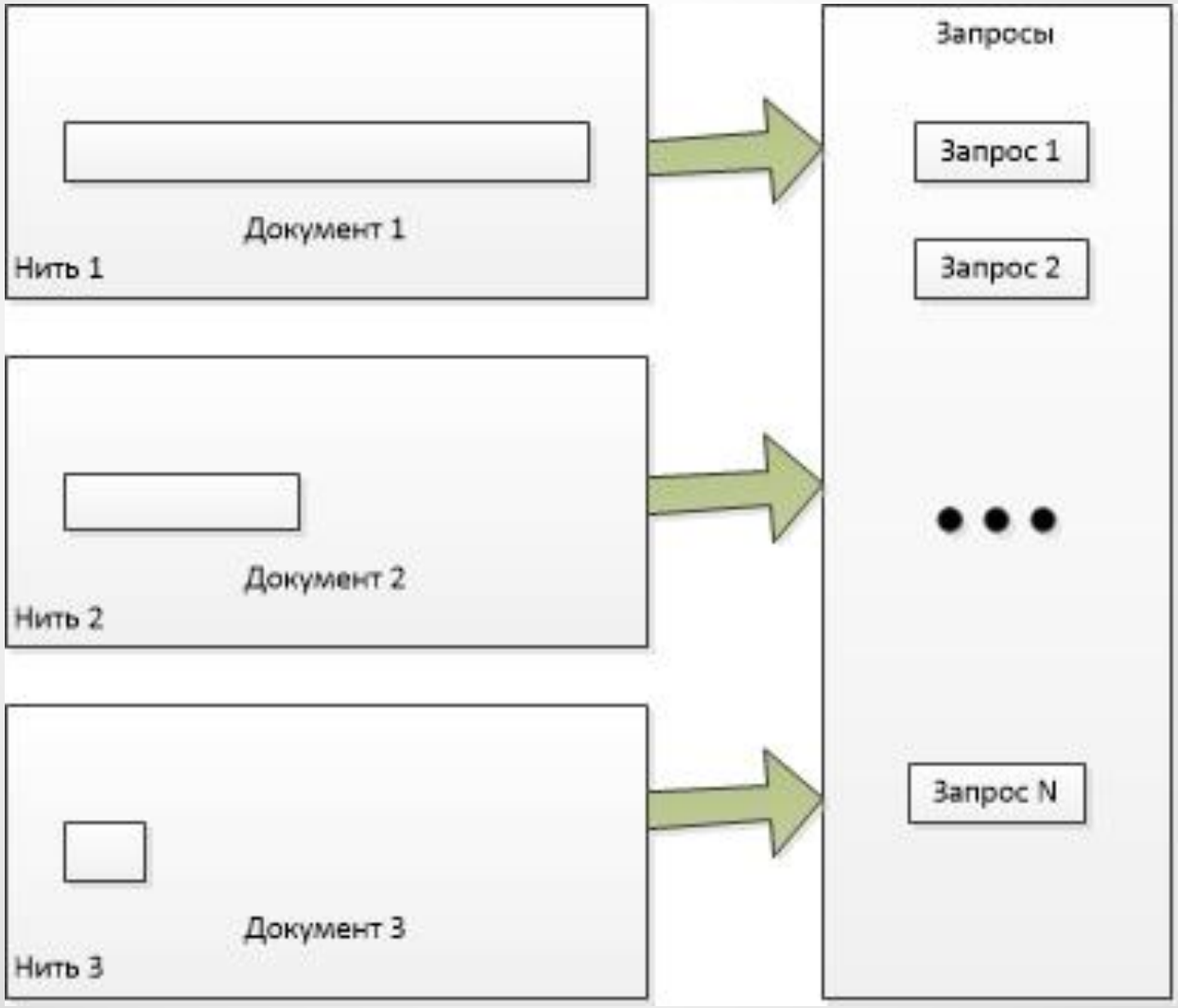


Разбиение на задачи

- Нить/документ
  - Проверяем все текущие запросы на соответствие в документе
  - Низкая гранулярность

- **Нить/уникальное слово**
  - Препроцессинг
  - Структура: уникальное слово в документе -> уникальные слова из запросов, которые ему соответствуют
  - Высокая гранулярность

Нить/документ



Нить/уникальное слово

Тесты

- Нить/документ - 28.3 с.
- Нить/уникальное слов - 9.9 с.

- Уменьшение количества вызовов дорогих функций
- Лучшее разбиение на задачи

Результаты

- Изучены архитектуры многопоточных приложений
- Реализовано 2 прототипа системы
- Результаты будут доложены на конференции ACM SIGMOD’2013