



**Нижегородский государственный университет
им. Н.И.Лобачевского**

Факультет Вычислительной математики и кибернетики

***“Технологии построения и использования кластерных
систем”***

Краткий обзор

Сенин А.В.
Кафедра математического
обеспечения ЭВМ

Содержание

- ❑ Введение
- ❑ Как подстроить кластер
- ❑ Программное обеспечение кластера
- ❑ Назначение систем управления кластерами
- ❑ Основные требования к системам управления
- ❑ Рассмотрение некоторых систем управления:
 - *Microsoft Compute Cluster Server 2003*
 - *Condor*
 - *Metaccluster*
- ❑ Планирование распределения ресурсов на кластерных системах:
 - *Backfilling*



Введение

Генная инженерия

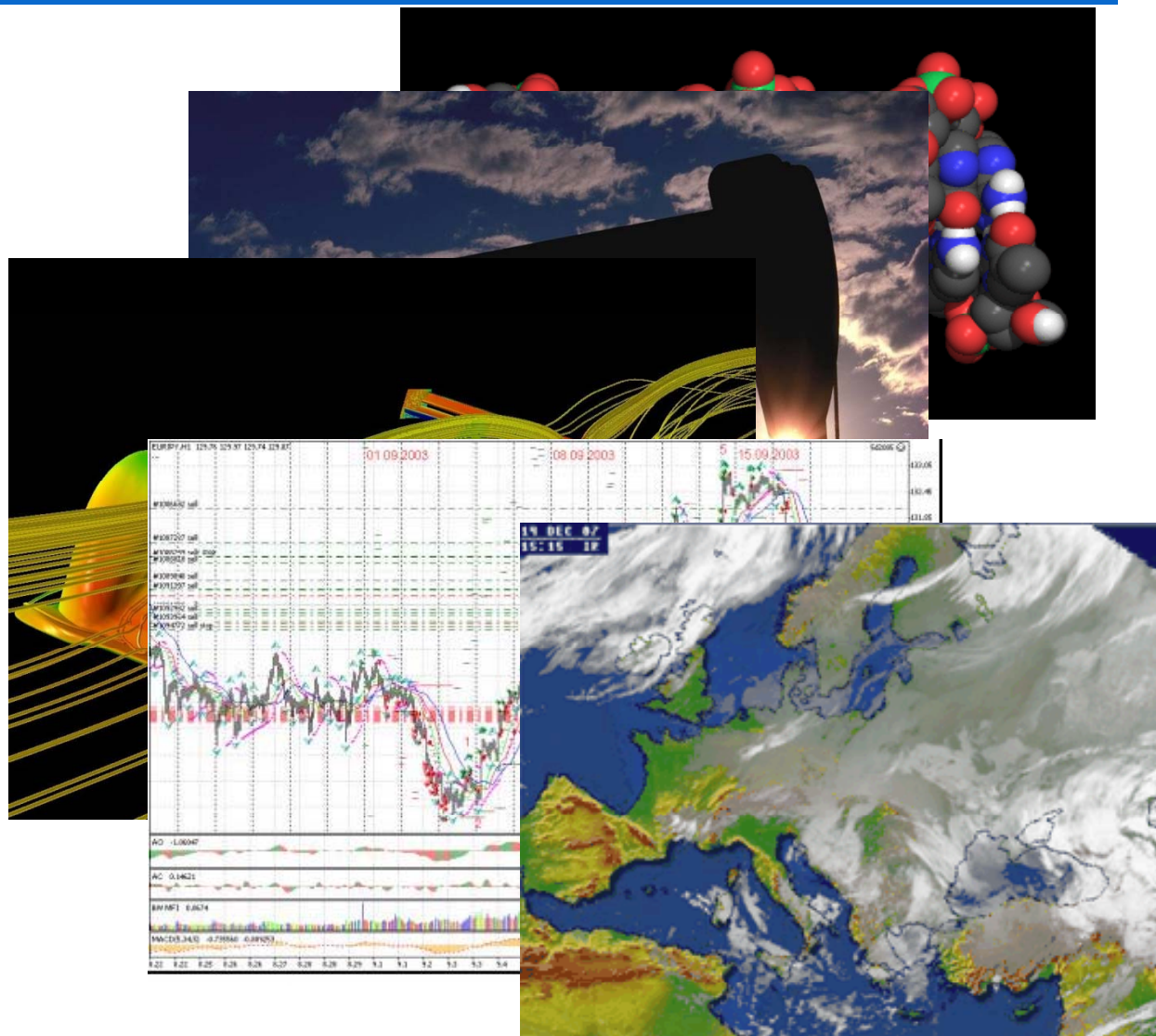
Поиск залежей нефти и газа

Физическое моделирование

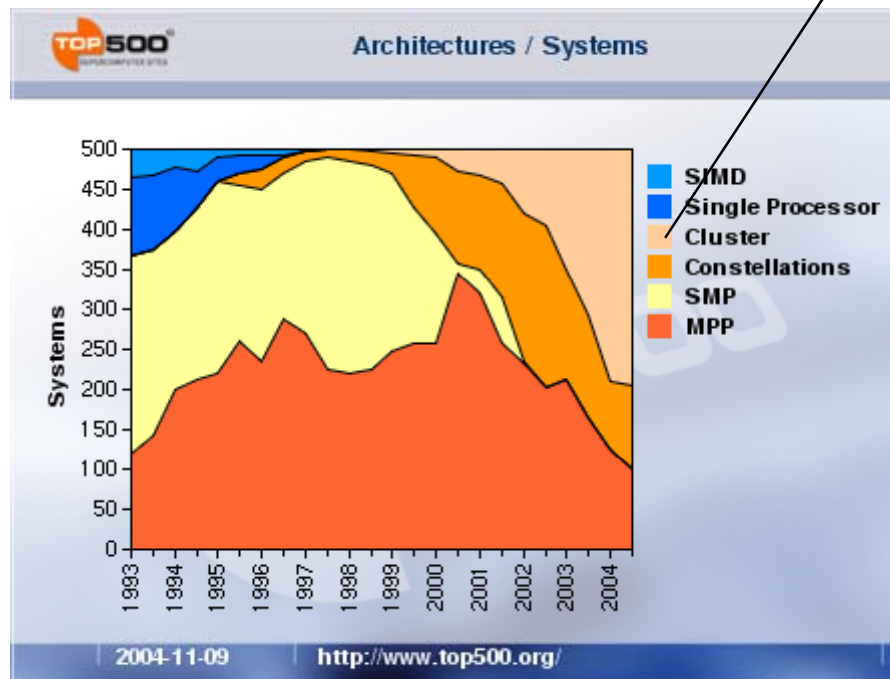
Финансовая аналитика

Прогнозирование погоды

...



Введение



В списке top 500 самых мощных компьютеров мира кластеры занимают около 75% (июль 2007)



Введение

- ❑ Параллельное программирование – основа высокопроизводительных вычислений.
- ❑ Кластер – множество отдельных компьютеров, объединенных в сеть. Ключевые особенности:
 - отсутствие общей памяти
 - пересылка данных по сети
- ❑ Кластеры стали фактическим стандартом высокопроизводительных систем. Основные достоинства:
 - дешевизна
 - простая расширяемость



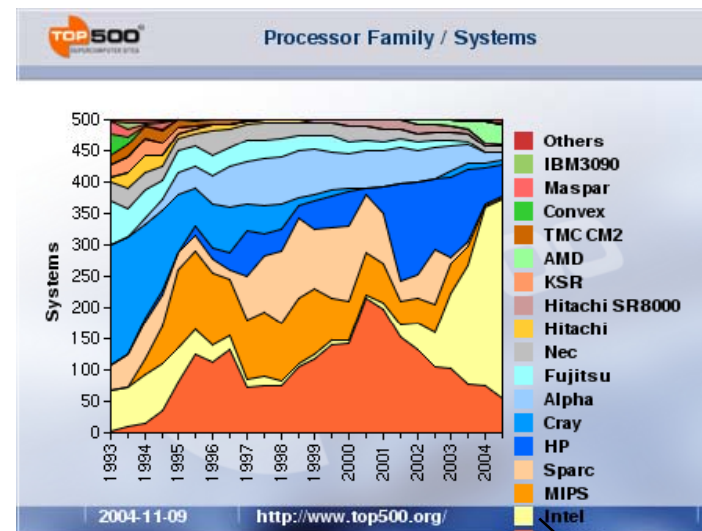
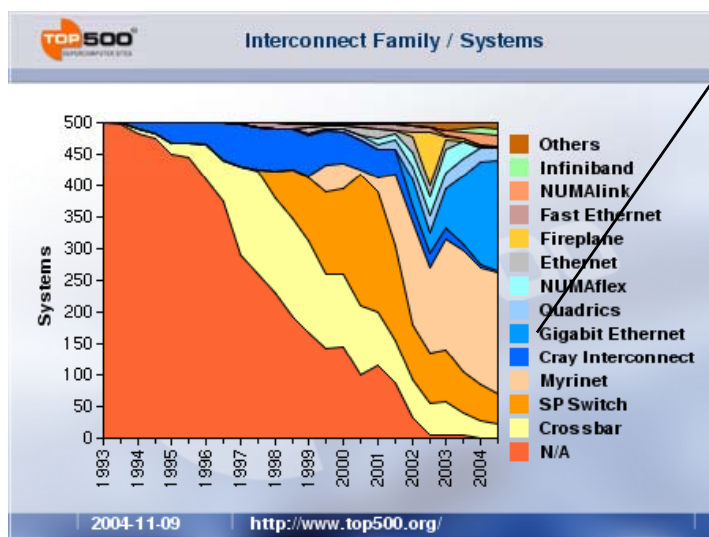
Введение

- ❑ Кластеры, как правило, используются одновременно большим числом пользователей.
- ❑ Вычислительные задачи очень требовательны к ресурсам и работают днями и даже неделями.
- ❑ Современная ситуация с кластерами очень напоминает ситуацию с мэйнфреймами на заре компьютерной эпохи.
- ❑ Необходимы специализированные системы, способные эффективно управлять ресурсами кластера.



Введение

Наиболее часто используемые сетевые интерфейсы:
Gigabit Ethernet,
Myrinet



Наиболее распространенные процессоры – Intel
Наиболее распространенное количество ядер - 2



Как построить кластер

- ❑ Обратиться в специальную компанию: "Т-платформы", Kraftway, IBM, HP, ...
или
- ❑ Собрать самому «кластер рабочих станций» (cluster of workstations):
 - Несколько компьютеров, купленных в обычном магазине, без монитора, без графической карты, часто без жесткого диска
 - Сеть (Gigabit Ethernet, Myrinet, InfiniBand...)
 - Качественный коммутатор
 - Выделенный управляющий компьютер, файловое хранилище
 - Система охлаждения

или

- ❑ Получить доступ к действующему кластеру



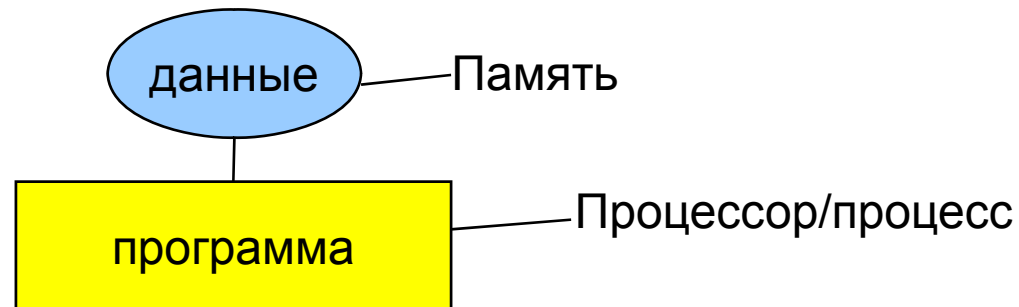
Программное обеспечение кластера

- ❑ В качестве операционной системы на кластерах используют различные версии ОС Unix и Linux (в сумме более 90% операционных систем списка top500)
- ❑ В последнее время увеличивается число кластеров на базе ОС Windows (6 установок в текущем списке top500)
- ❑ MPI – наиболее распространенная коммуникационная библиотека

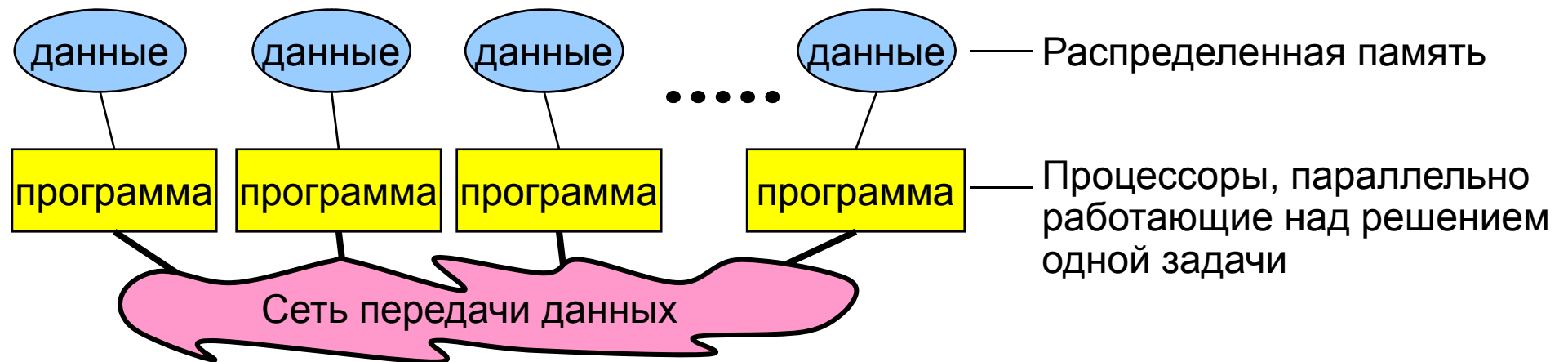


Программное обеспечение кластера

- ❑ Парадигма последовательной программы



- ❑ Парадигма параллельной MPI программы



Программное обеспечение кластера

- ❑ Помимо интерфейса передачи сообщений необходимо специально ПО для управления потоком задач
- ❑ Два основных подхода к управлению кластерами:
 1. Распределенные окружения для кластерных вычислений
 2. Системы управления кластерами
- ❑ Идея первого подхода состоит в создании «кластерной» операционной системы
 - + Подобные среды весьма удобны, работа в них фактически не отличается от работы в обычной операционной системе
 - Необходима коррекция операционной системы. Существуют реализации только для UNIX-подобных операционных систем.
- ❑ Системы управления кластерами работают поверх операционной системы. Они являются развитием систем пакетной обработки. Более распространенная технология



Назначение систем управления кластерами

- ❑ Автоматизация процессов
 - прием задач от пользователей
 - постановка в очередь
 - планирование и запуск
 - возврат результатов вычислений
- ❑ Предоставление удаленного доступа к кластеру
- ❑ Обеспечение максимальной пропускной способности кластера путем эффективного планирование запуска.



Основные требования

- ☐ Производительность
- ☐ Надежность
- ☐ Удобство использования



Основные требования

Производительность

❑ Эффективное планирование

- Минимизация времени выполнения задач
- Минимизация среднего времени ожидания задачи в очереди

❑ **Перепланирование** – система управления должны отслеживать состояние заданий и загрузку узлов кластера, чтобы при необходимости перераспределить пользовательские процессы (например, в случае отказа одного из узлов)



Основные требования

Надежность

- ❑ **Устойчивость к сбоям** – система должна оперативно обрабатывать возникающие аппаратные и программные сбои
- ❑ **Контрольные точки** – сохранение состояния выполняемого процесса через регулярные промежутки времени, чтобы в случае аварийного выхода узла из строя не начинать вычисления сначала
- ❑ **Безопасность** – весь спектр вопросов, связанных с защищенностью системы от возможных атак: безопасные протоколы общения компонентов системы, аутентификация пользователей



Основные требования

Удобство использования

- ❑ **Множественность интерфейсов** – наличие нескольких интерфейсов доступа, чтобы пользователь мог выбрать наиболее удобный: web-интерфейс, графический интерфейс (окно Windows/Linux), командный интерфейс, программный интерфейс
- ❑ **Конфигурируемость** – административный интерфейс с возможностью выбора алгоритма планирования, добавления/удаления пользователей, изменения прав без остановки счета на кластере
- ❑ **Информативность** – предоставление информации о состоянии заданий и общей статистики о запусках и загрузке узлов кластера



Обзор систем управления, Condor

- ❑ Система управления кластером Condor была создана группой разработчиков университета Висконсин-Мэдисон (Wisconsin-Madison). Первая конфигурация была развернута в этом университете в 1986 году
- ❑ Ключевые особенности системы:
 - Создание контрольных точек с возможностью миграции заданий между процессами
 - Возможность интеграции в Grid
 - Поддержка широкого круга архитектур и операционных систем
 - Использование для вычислений не только выделенных машин, но и рабочих станций



Condor
High Throughput Computing



Обзор систем управления, Condor

□ Типы узлов в системе:

- **Центральный менеджер.** Ядро системы. Собирает информацию о загрузке узлов кластера и получает запросы пользователей, он затем определяет время и место запуска. Остановка менеджера приведет к остановке всей системы
- **Выполняющие машины.** Счетные узлы. Производительность выполняющих машин в конечном итоге определяет, когда пользователь получит результат. Однако, Condor создавался с намерением использовать все доступные ресурсы, в том числе и низкопроизводительные



Обзор систем управления, Condor

□ Типы узлов в системе:

- **Запускающие машины.** Узлы, с которых производится запуск пользовательских заданий. Во время работы задания на запускающей машине будут выполняться некоторые служебные процессы, которые обеспечивают доступ удаленных процессов к локальным файлам
- **Сервер контрольных точек.** Узел для хранения всех контрольных точек пользовательских заданий. От узла требуется большой объем оперативной памяти и надежное сетевое соединение



Обзор систем управления, Microsoft CCS

- ❑ Сравнительно новая система управления кластером (9 июня 2006 года – дата выхода первого релиза)
- ❑ Первая система управления, предназначенная исключительно для Windows
- ❑ Ключевые особенности системы:
 - Наличие параллельного отладчика MPI – программ
 - Автоматическое освобождение ресурсов при аварийной остановке задания
 - Интеграция с средствами администрирования Windows Server: Active Directory, Microsoft Operations Manager 2005, Systems Management Server 2003, MMC 3.0



Обзор систем управления, Microsoft CCS

- ❑ Управление может осуществляться через графический (основной), командный, программный интерфейсы

The screenshot displays two overlapping windows from the Microsoft CCS (Cluster Control System) management tools.

The top window, titled "Job Queue at s-cw-head", shows a table of submitted jobs:

ID	Name	Priority	Submitted By	Status	Submit Time	Pending Reason
1	hoothname	Normal	CCAM\Senin	Finished	21.06.2006 4:10:14	
2	imb	Normal	CCAM\Senin	Failed	21.06.2006 4:19:50	
3	imb	Normal	CCAM\Senin	Failed	21.06.2006 4:24:24	
4	imb	Normal	CCAM\Senin	Failed	21.06.2006 4:25:23	
5	imb	Normal	CCAM\Senin	Finished	21.06.2006 5:13:20	

The bottom window, titled "s-cw-head - Remote Desktop" and "Compute Cluster Administrator", shows a list of cluster nodes:

Name	Status	Jobs Ru...	CPUs	CPUs in...	OS Vers...	Total M...
S-CW2-01	Ready	0	4	0	5.2.3790	2046
S-CW2-02	Ready	0	4	0	5.2.3790	2046
S-CW2-03	Ready	0	4	0	5.2.3790	2046
S-CW2-04	Ready	0	4	0	5.2.3790	2046
S-CW2-05	Ready	0	4	0	5.2.3790	2046
S-CW2-07	Ready	0	4	0	5.2.3790	2046
S-CW2-09	Ready	0	4	0	5.2.3790	2046
S-CW2-10	Ready	0	4	0	5.2.3790	2046
S-CW2-11	Ready	0	4	0	5.2.3790	2046
S-CW2-12	Ready	0	4	0	5.2.3790	2046
S-CW2-13	Ready	0	4	0	5.2.3790	2046
S-CW2-14	Ready	0	4	0	5.2.3790	8190
S-CW2-15	Ready	0	4	0	5.2.3790	8190
S-CW-HEAD	Ready	0	4	0	5.2.3790	2046
S-CW2-08	Pending Approval	0	4	0	5.2.3790	2046

The "Compute Cluster Administrator" window also includes a sidebar with "Node Management" actions (Add Nodes, Manage Cluster Users, Remote Desktop Settings, Launch Compute Cluster ...) and a "Select one node in the list to view active jobs." prompt.



Обзор систем управления, Метакластер

- ❑ Разрабатывается в ННГУ
- ❑ В настоящее время используется на кластере ННГУ
- ❑ Ключевые особенности системы:
 - Возможность интеграции нескольких кластеров под единым управлением
 - Интеграция с системами управления сторонних разработчиков
 - Web – интерфейс доступа, программный интерфейс

Нижегородский Государственный Университет им Н.И.Лобачевского
Институт информационных технологий и системного обеспечения ЭВМ

УПРАВЛЕНИЕ КЛАСТЕРОМ

Постановка задания в очередь:

Имя задачи:	
Файл результатов:	
Количество процессов:	
Параметры запуска:	
Кластер:	NNSU Cluster (MS CCS 2003 based) ▼
Режим работы:	MS MPI ▼
Запустить	Отмена



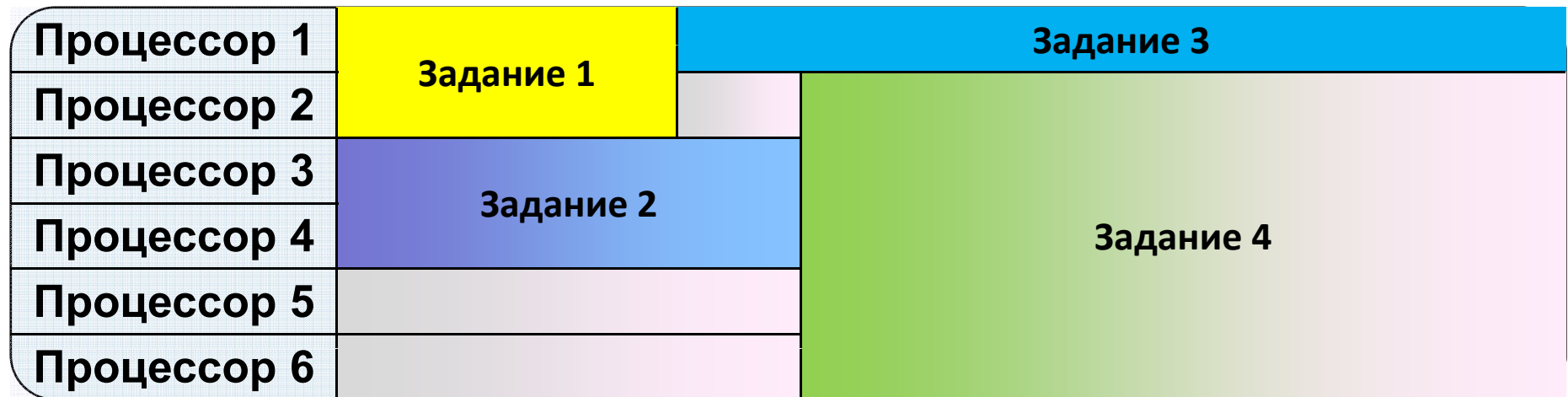
Планирование распределения ресурсов на кластерных системах

- ❑ Эффективное распределение ресурсов – важная задача при использовании кластера. Критерии эффективности:
 - Минимизация времени выполнения задач
 - Минимизация среднего времени ожидания задачи в очереди
- ❑ Классическая задача составления расписания NP – полна, поэтому на практике используются приближенные методы или эвристики
- ❑ Использование точного метода составления расписания осложняется еще и тем фактом, что задачи в систему могут поступать. И полное перепланирование пришлось бы производить для каждой новой задачи
- ❑ Существует специализированные планировщики для кластеров: Maui, Moab. Обе системы имеют возможность интегрироваться с системами управления кластеров сторонних производителей



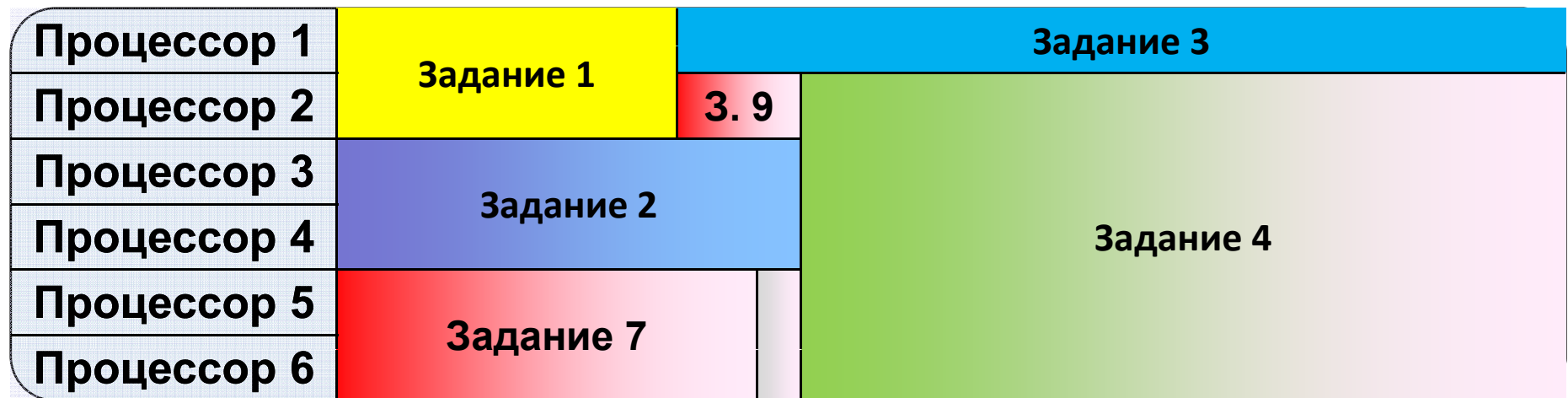
Backfilling

- ❑ Первая стадия алгоритма обратного заполнения – составление расписания для наиболее приоритетных N задач с учетом пользовательской оценки требуемого времени
- ❑ В расписании образуются окна. В эти периоды процессоры ожидают начала выполнения запланированных заданий



Backfilling

- ❑ Вторая стадия алгоритма – заполнение образовавшихся окон менее приоритетными и менее требовательными к ресурсам задачами
- ❑ Таким образом, алгоритм Backfilling старается обеспечить максимальную загрузку кластера в каждый момент времени



Backfilling

- ❑ При подсчете приоритетов может использоваться следующая информация:
 - Приоритет пользователя (например, преподаватель и студент)
 - Приоритет, который пользователь поставил своей задаче
 - Время ожидания запуска
 - Число задач, запущенных пользователем за последний день/неделю/месяц
 - Предположительное время выполнения задачи
 - ...



Вопросы...

- ❑ Вопросы, замечания, дополнения...

- ❑ Контактная информация:
 - Сенин Андрей
 - Аспирант факультета ВМК, каф. МО ЭВМ
 - andrey.senin@itlab.unn.ru

