

Методы анализа динамических характеристик работы суперкомпьютера

Никитенко Дмитрий Александрович

*научный сотрудник
НИВЦ МГУ имени М.В. Ломоносова*

семинар кафедры математического обеспечения ЭВМ
факультета вычислительной математики и кибернетики
ННГУ им. Н.И.Лобачевского

Научный руководитель: чл.-корр. РАН, проф. Воеводин Вл.В.

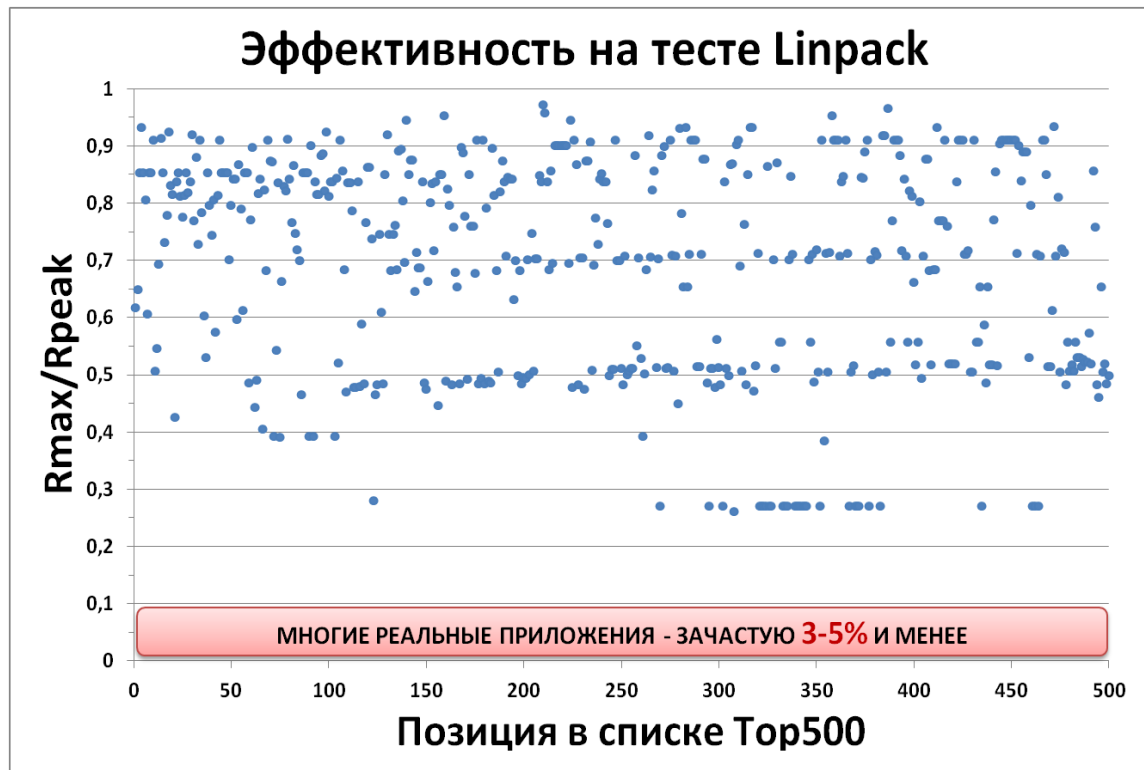
Работа выполнена в НИВЦ МГУ имени М.В. Ломоносова

План доклада

- Актуальность исследований
- Анализ существующих решений
- Постановка задачи
- Описание и апробация предложенных методов
- Основные результаты работы

Низкая эффективность суперкомпьютерных приложений и систем

Эффективность –
отношение реальной
производительности
к пиковой
($Eff = R_{max} / R_{peak}$)

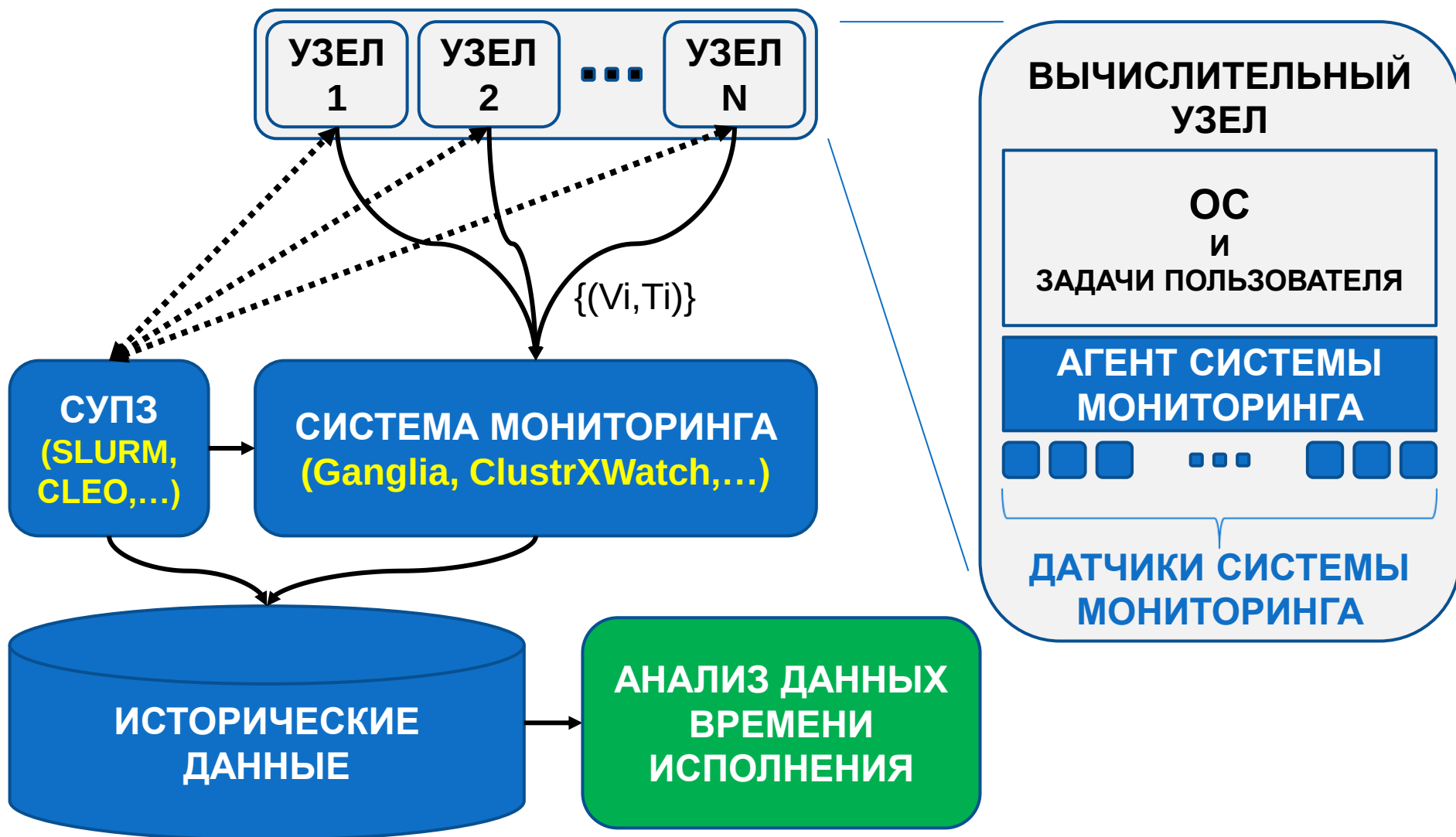


Проблема: эффективность работы отдельных реальных суперкомпьютерных приложений и, как следствие, суперкомпьютерных систем зачастую крайне мала!

Существующие подходы и инструменты, направленные на анализ производительности приложений

- Анализ трасс
 - Scalasca, Score-P, Vampir, ThreadSpotter, HPCToolkit, ...
 - Профилирование
 - VTune, DVM, ParaDyn, PerfMiner, ...
 - Статические и другие методы (анализ корректности кода и т.д.)
 - V-Ray, Intel Inspector XE, ...
- Направлены на глубокое исследования свойств отдельных программ
- Требуют подготовительных работ
- Не могут быть применены для анализа произвольной задачи из потока

Использование данных системного мониторинга и СУПЗ



Постановка задачи

Разработка комплексного подхода к исследованию причин и признаков снижения эффективности суперкомпьютерных приложений и систем на основе анализа потока данных системного мониторинга.

Комплексный подход должен учитывать неоднородность вычислительной системы и соответствующее деление счетного поля на разделы. Комплексный подход должен быть направлен на нахождение истинных причин снижения эффективности суперкомпьютерных приложений и систем и на анализ соответствия свойств разделов суперкомпьютера и суперкомпьютерных приложений.

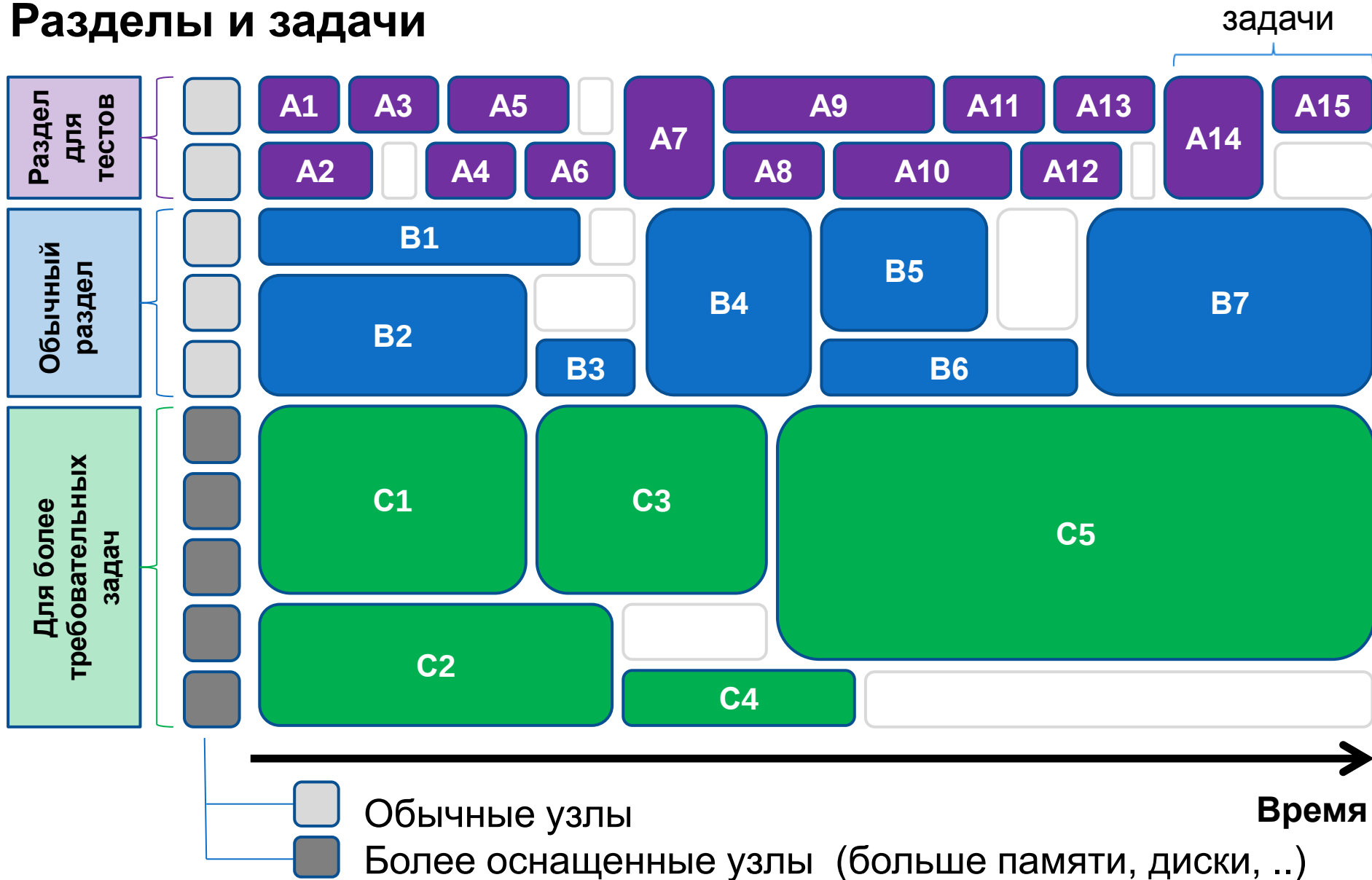
Комплексный подход к исследованию эффективности суперкомпьютерных приложений и систем

Три уровня исследования эффективности:

- **эффективность суперкомпьютерных приложений**
- **соответствие свойств приложений разделам суперкомпьютера**
- **эффективность использования ресурсов системы в целом**

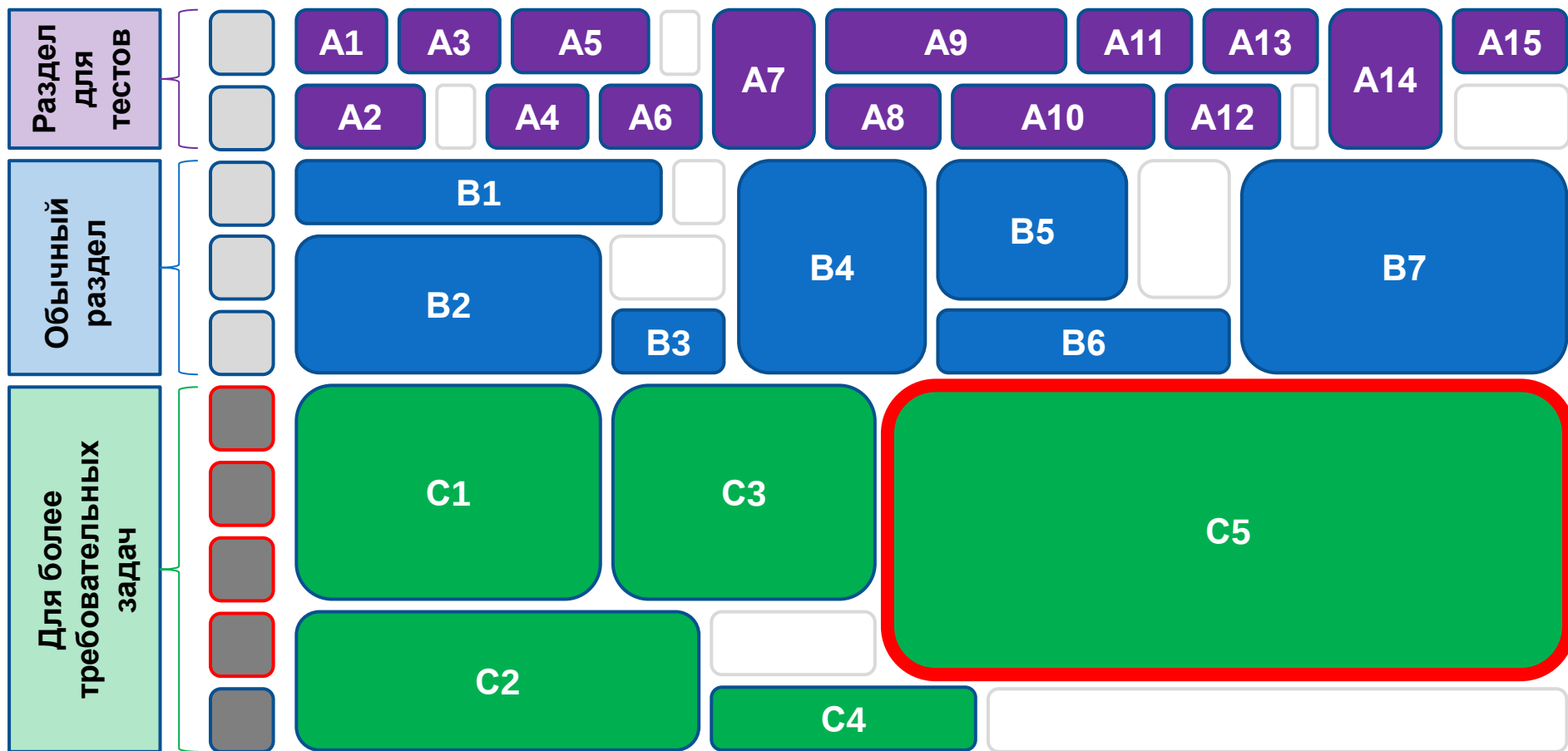
Особенности целевых суперкомпьютерных сред

Разделы и задачи



Эффективность суперкомпьютерных приложений

задачи

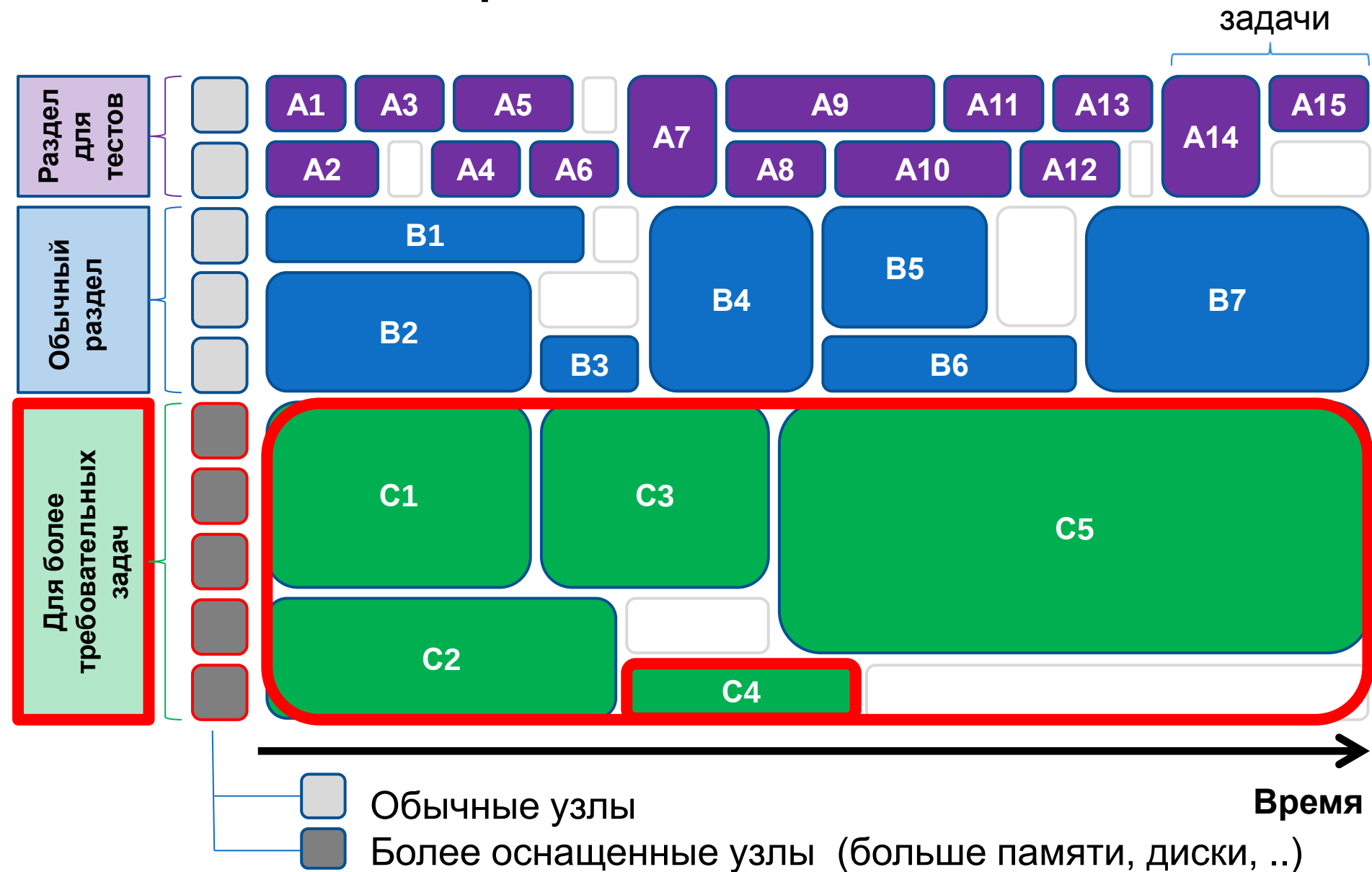


Обычные узлы

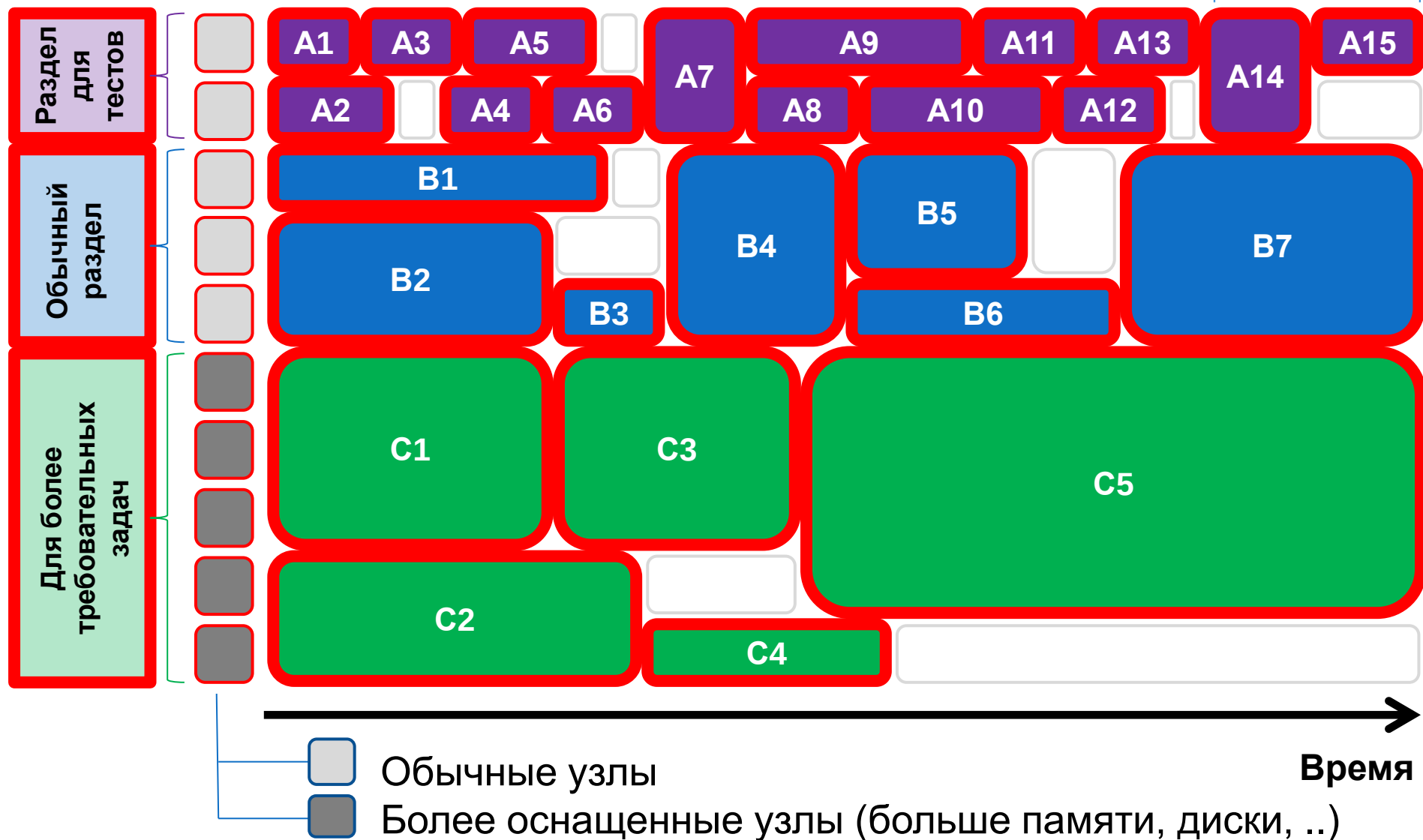
Более оснащенные узлы (больше памяти, диски, ..)

Время

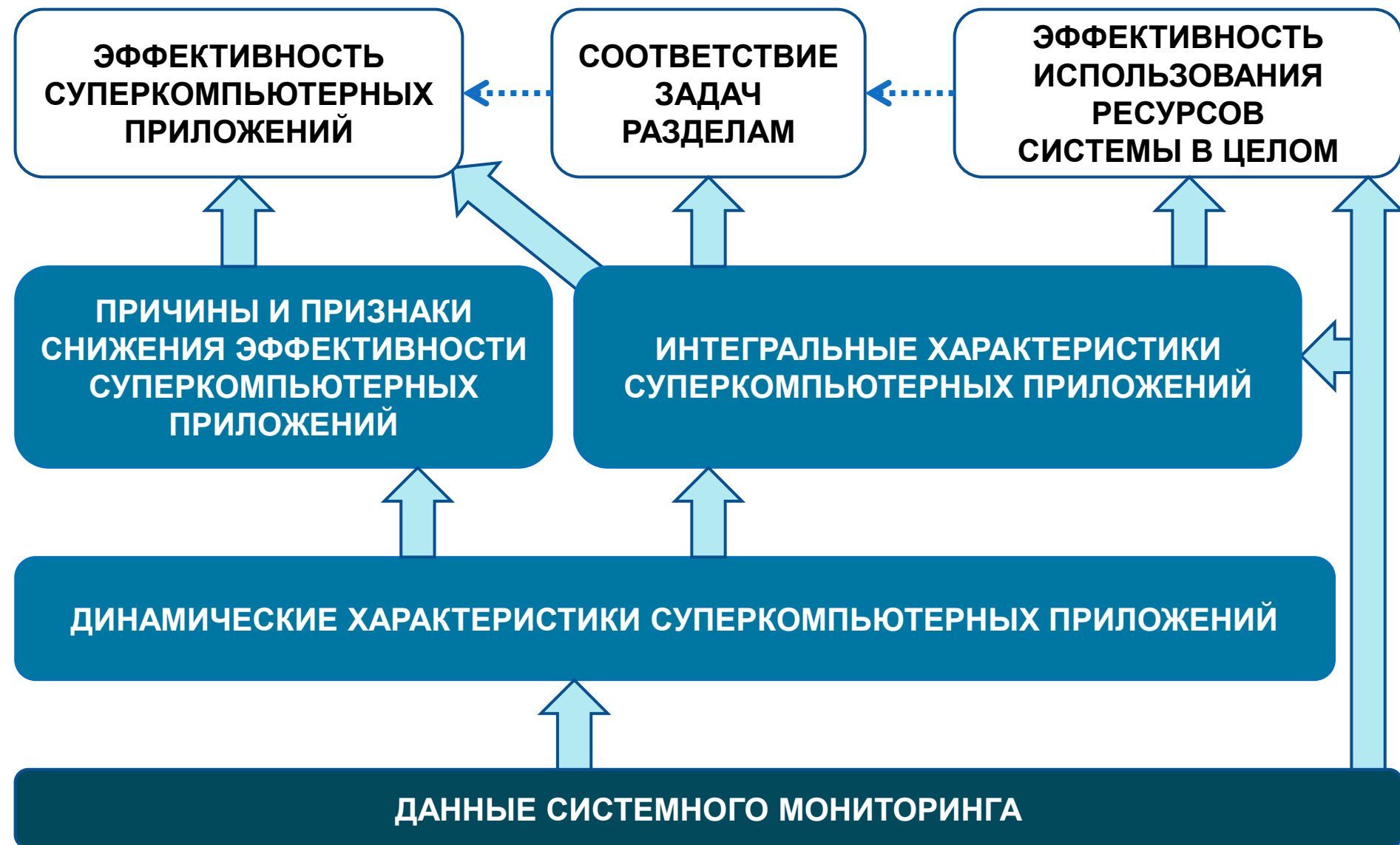
Соответствие задач разделам



Эффективность использования ресурсов системы в целом задачи

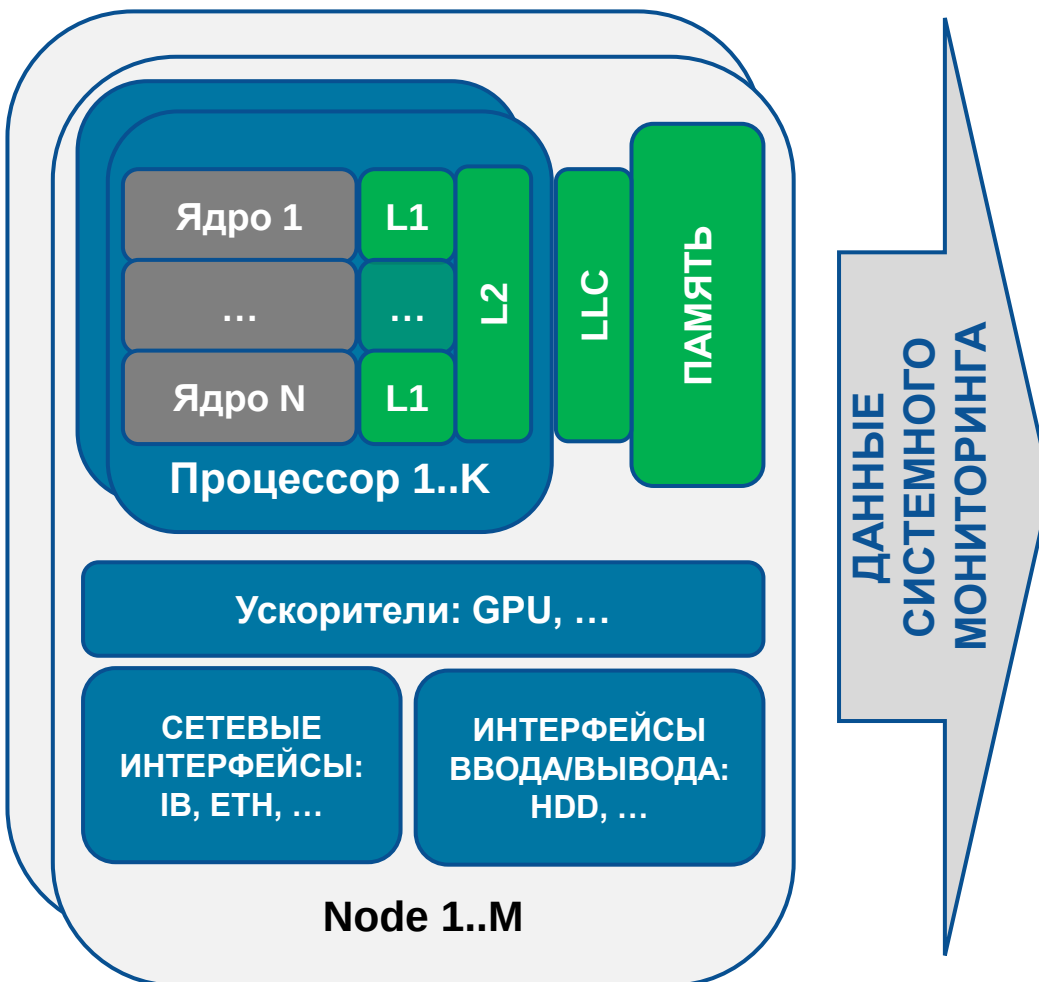


Системный мониторинг как единая основа подхода



Ключевой набор динамических характеристик

- Отражает степень использования основных ресурсов приложением
- Состав и частота съема определяются исходя из специфики исследований



Динамические характеристики:

Загрузка процессора:

- пользовательская (CPU user)
- число операций с плавающей точкой (Flop/s)
- число процессов, готовых принять управление (LoadAverage)

Работа с памятью:

- свопинг
- число обращений к памяти
- число промахов в кэш
- объем свободной памяти

Ввод/вывод:

- интенсивность работы с ФС (скорость чтения и записи)

Коммуникации:

- интенсивность межзвлового обмена (скорость чтения и записи)

Объем данных системного мониторинга системах

Исходные данные:

одно значение в секунду от агентов с узлов
(~10 Тбайт сырых данных в день с
системы масштаба «Ломоносова»)

Фильтрация перед сохранением:

или изменение более, чем на 10% (20% для
процессорных)
или не посылались последние 30 секунд

Корректность интерпретации источников данных при построении и анализе динамических характеристик

Определение типовых значений и области их изменения

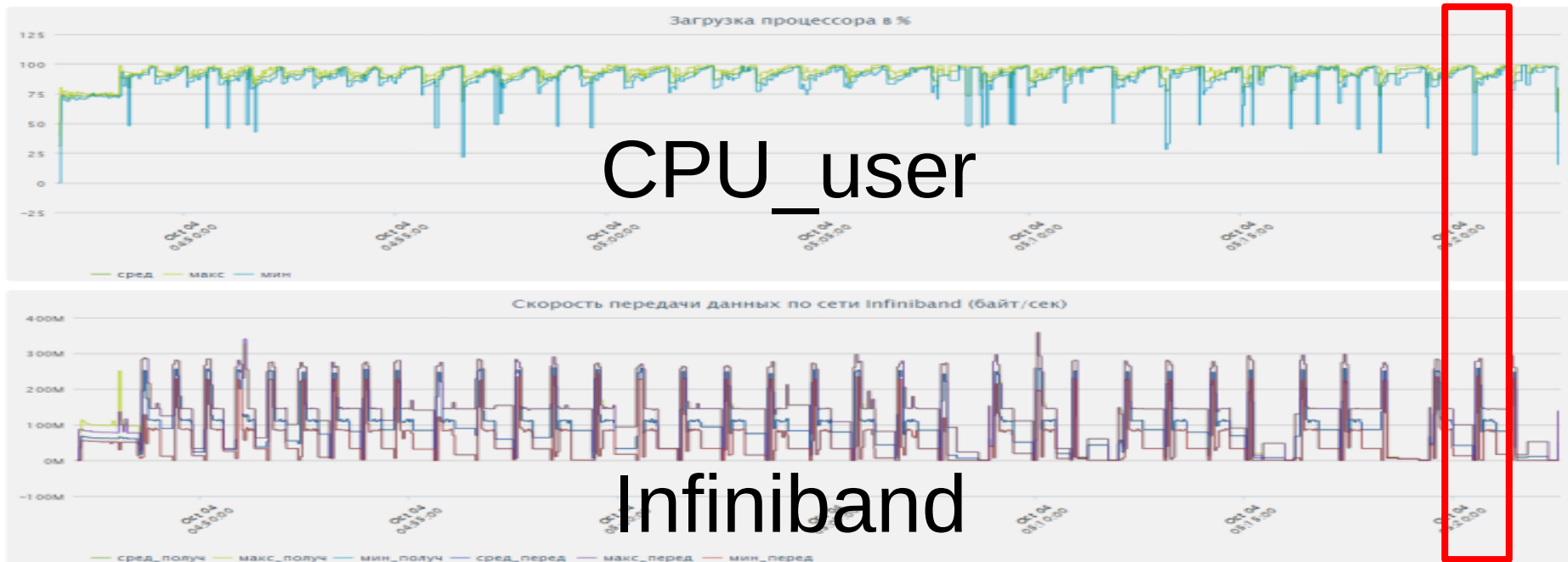
Значение	Трактовка
0	Нет готовых к взятию управления процессов
$N \cdot K$	Полная загрузка процессоров при отсутствии конкуренции процессов. Идеальный вариант.
На $1..K$ ниже $N \cdot K$	Высокая, но не полная загрузка
$N \cdot K < x < 1,5 N \cdot K$ для N -ядерного процессора	Повышенная конкуренция между процессами
Много более $N \cdot K$	Очень большая конкуренция. Процессов больше, чем ресурсов, и часть всегда будет простаивать. Частая смена контекстов.
Менее 0	Не соответствует природе характеристики

Корректность интерпретации источников данных при построении и анализе динамических характеристик

Детальное описание источника данных

Поле	Описание
name	Имя датчика
id	Идентификатор датчика
description	Понятное человеку описание назначения датчика
unit	Единицы измерения «bytes/sec», «%» и т.д.
val_type	Тип значения, абсолютный или накопительный - {absolute; accumulated}
scope	Когда измерено значение {now; sinceLast; untilNext}
supported	Поддерживается ли в данной реализации системы мониторинга
avail	Доступность в данный момент (м.б. отключена временно)
source	Источник {clustrxwatch, ganglia, и т.д.}
granularity	Частота съема в Гц
min	Разумный минимум области допустимых значений
max	Разумный максимум области допустимых значений
nodes	Список узлов, поддерживающих данную метрику

Взаимосвязанное поведение динамических характеристик



Вытекает из исходной взаимосвязи ресурсов, характер использования которых они отражают

Идея выявления причин и признаков снижения эффективности

Причины и признаки снижения производительности суперкомпьютерных приложений

- Признак – аномальное, необычное поведение приложения, сигнализирующее о снижении эффективности работы
 - снижение уровня загрузки процессора
 - рост числа промахов при обращении в кэш
 - число процессов, готовых принять управление, меньше числа вычислительных ядер
 - ...
- Причина – обстоятельство, вызвавшее это поведение
 - интенсивный ввод-вывод
 - разбалансировка загрузки
 - низкая локальность данных
 - ...

Признаки снижения производительности суперкомпьютерных приложений

Выявление признаков неэффективного поведения опирается на выявление изменений в характере поведения соответствующих данному ресурсу динамических характеристик

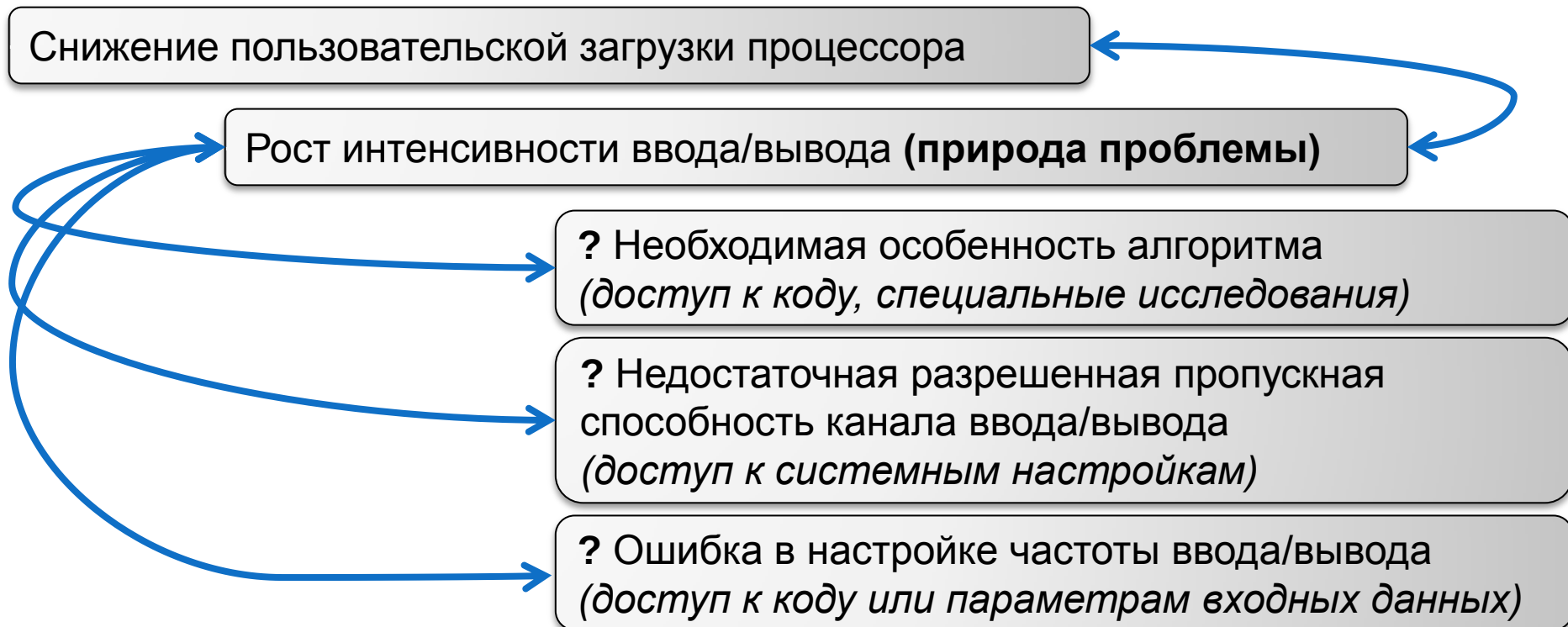
Вид поведения	Характер изменений в поведении
"полка"	неизменный (с погрешностью) уровень последних значений
рост	увеличение значений (быстрый рост и т.п.)
падение	уменьшение значений (быстрое падение и т.п.)
максимум	преодоление некоторого максимально возможного порога
минимум	преодоление некоторого минимально возможного порога
всплеск	временное существенное отклонение от "полки"
скачок	изменение "полки"
"шум"	не выявленная закономерность изменений значений
"осцилляция"	частое изменение, скачки значений на коротком интервале времени

Природа причин снижения производительности суперкомпьютерных приложений

- **Использование процессора:**
 - не используются ресурс параллелизма: векторизация, конвейеризация и т.п.;
 - не задействованы специальные оптимизирующие возможности процессора;
 - простои на ожидании приема сообщений;
 - простои из-за избыточных синхронизаций;
 - разбалансированность вычислительной нагрузки;
 - ...
- **Работа с памятью:**
 - низкая локальность вычислений, промахи в кэш;
 - утечка памяти;
 - свопинг;
 - ...
- **Ввод/вывод:**
 - ввод/вывод занимает значительную часть задачи;
 - избыточная работы с сетевой файловой системой;
 - создание необязательных временных файлов;
 - ...
- **Межузловой обмен:**
 - слишком частый обмен;
 - избыточный объем передаваемых данных;
 - ошибки интерфейсов;
 - низкое качество соединений;
 - ...

Причины и признаки снижения производительности суперкомпьютерных приложений

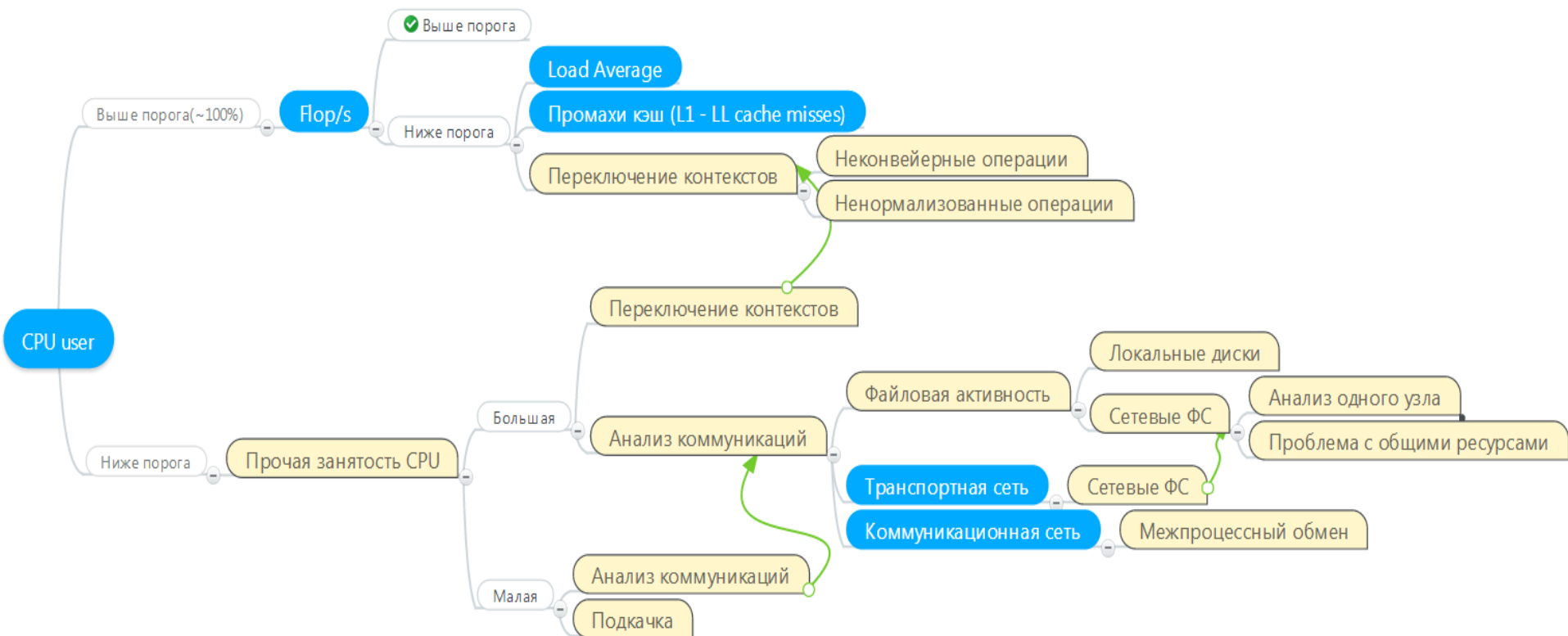
- Иерархия причин и признаков
- Определение природы возникновения проблемы
- Возможная ограниченность глубины анализа (недоступность кода, отсутствие информации о настройке системы и т.п.)



Общая схема метода анализа динамики выполнения суперкомпьютерных приложений

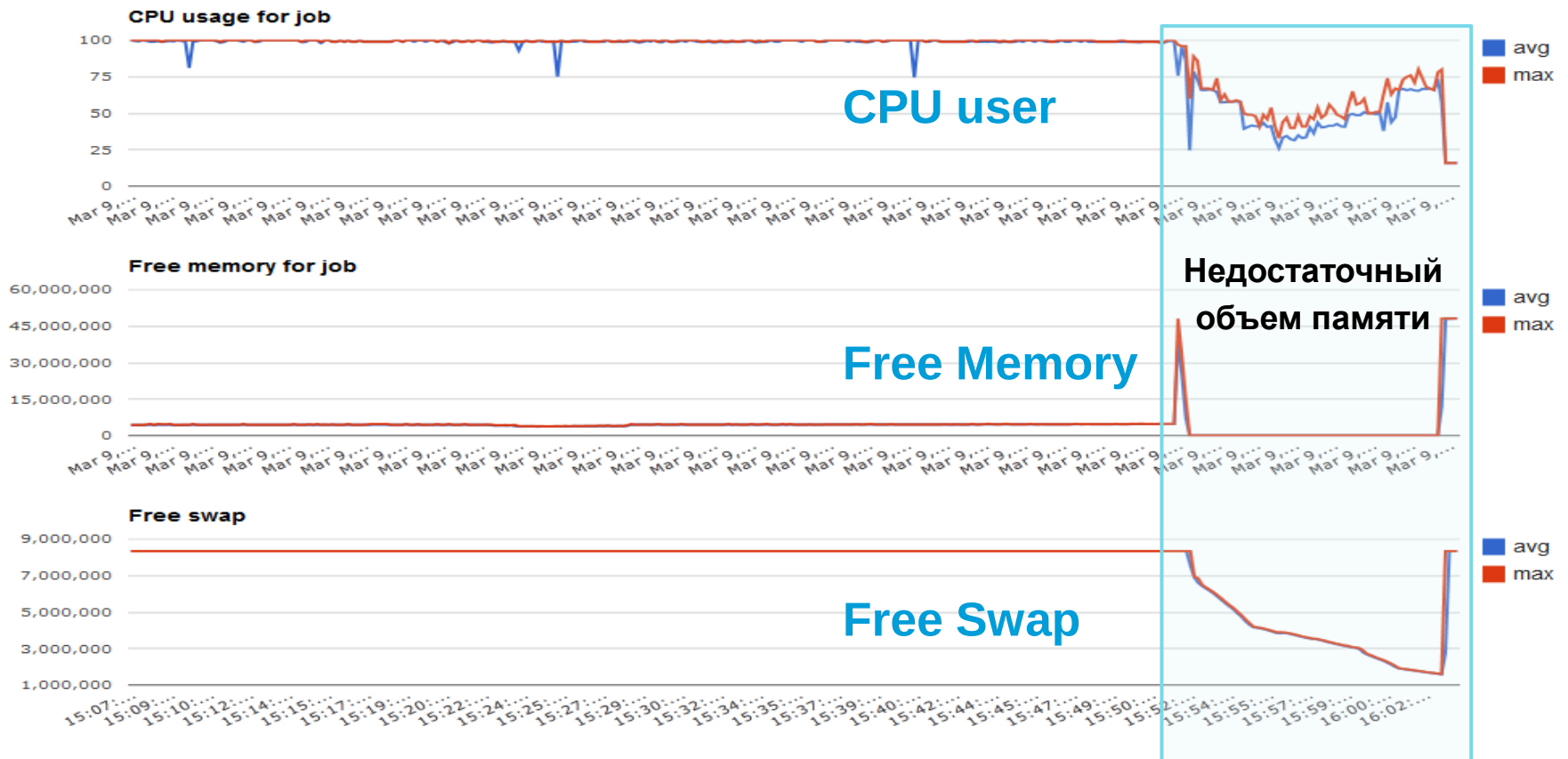


Пример последовательности исследования динамических характеристик суперкомпьютерных приложений



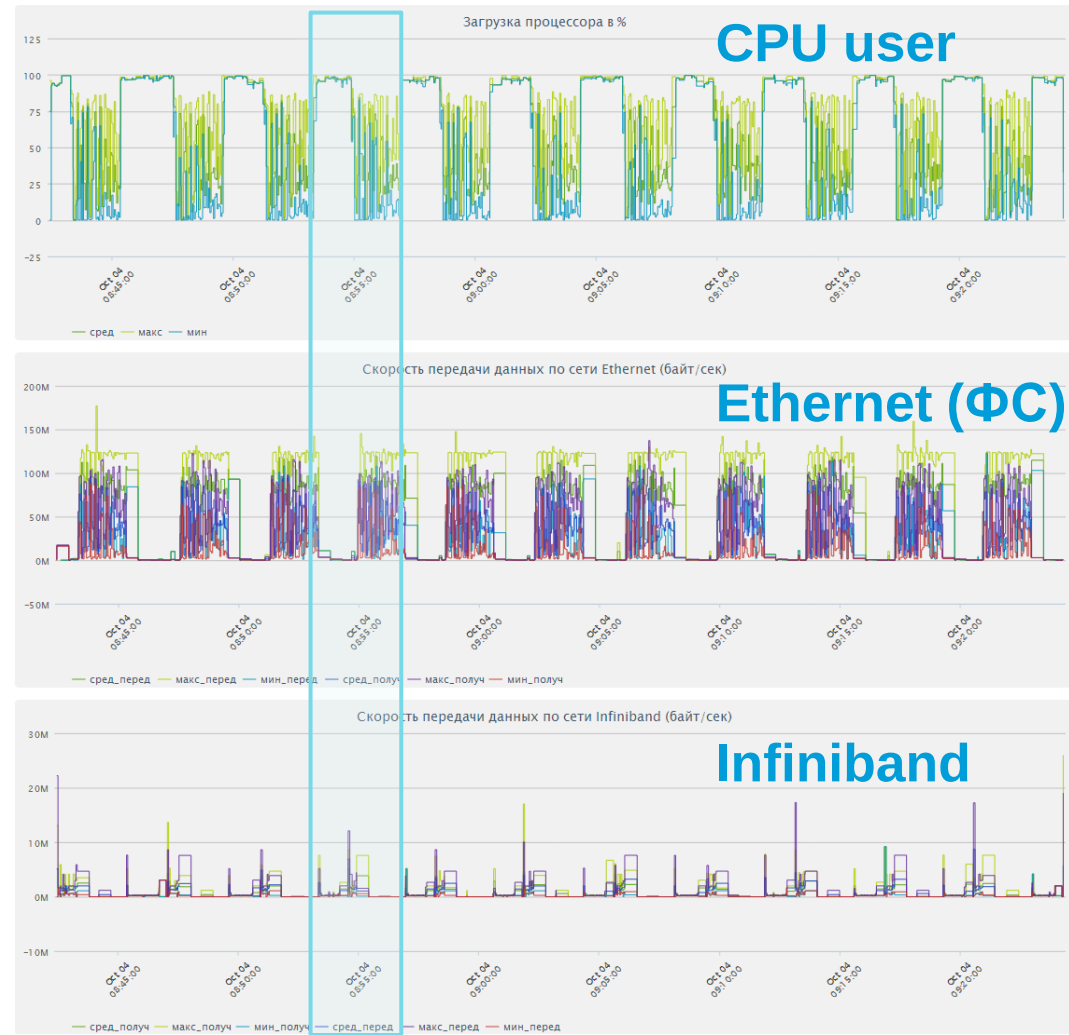
Анализ динамики выполнения суперкомпьютерных приложений

Nodes: cn11 CPU count: 3
Started: Fri Mar 9 15:07:09 2012 Queued: Fri Mar 9 15:07:09 2012
Ended: Fri Mar 9 16:04:04 2012 Wait time: 0 days 00:00:00
CPU*Hours: 0 Run time: 0 days 00:56:55



Анализ динамики выполнения суперкомпьютерных приложений

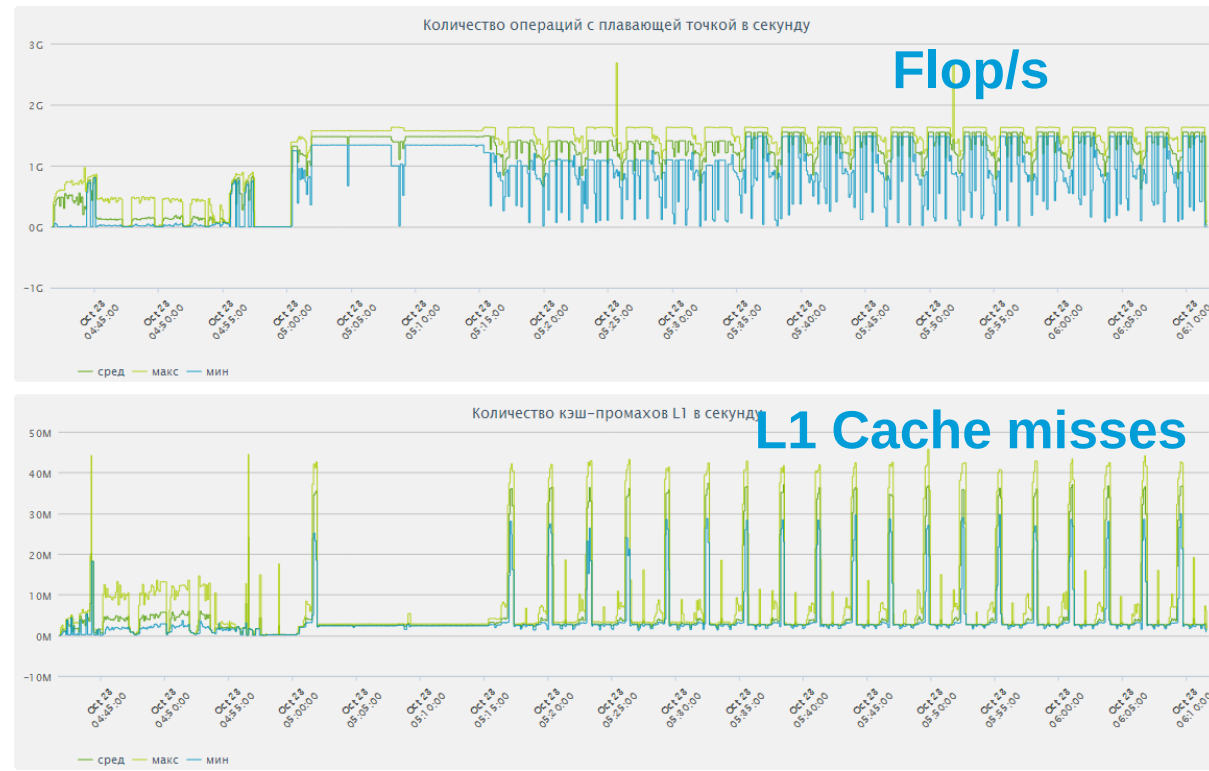
- Счет и обмен с ФС почти одинаковое время
- Обмен малыми порциями?



Анализ динамики выполнения суперкомпьютерных приложений

Локальность данных и вычислений

- Множественные промахи в кэш и снижение Flop/s



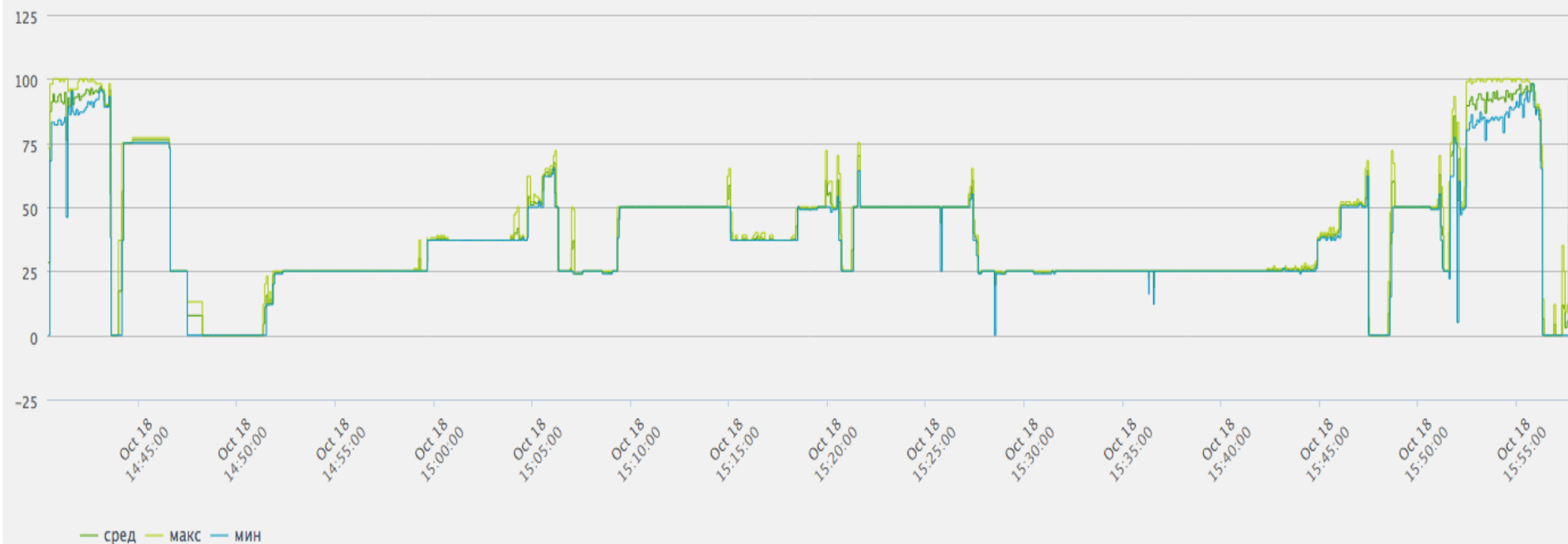
Анализ динамики выполнения суперкомпьютерных приложений

CPU user

Загрузка процессора в %

analyze_cpu_tpl8_hdd-1382104736-95272.csv 00:02:16.849 (18.656 lines/sec) transform

Загрузка процессора в %



Дисбаланс загрузки

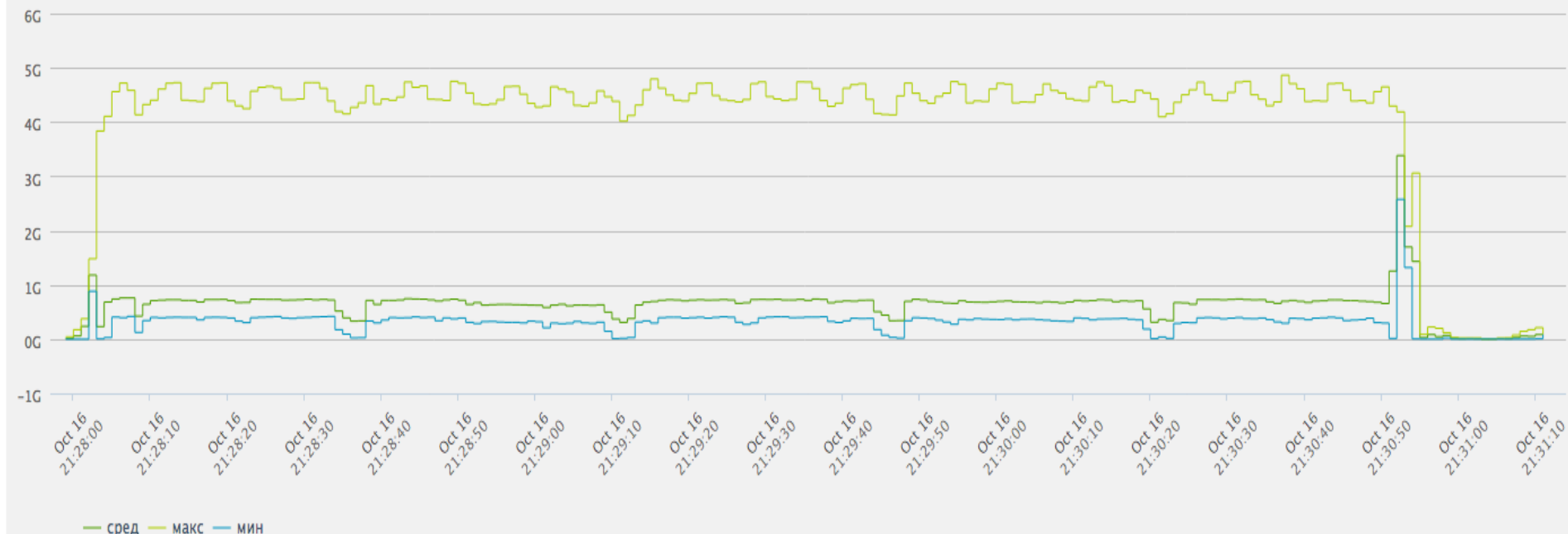
Анализ динамики выполнения суперкомпьютерных приложений

Flop/s

Количество операций с плавающей точкой в секунду

analyze_cpu_flops_tpl8_test-1381958878-32375.csv 00:00:36.200 (16.077 lines/sec) transform

Количество операций с плавающей точкой в секунду



Дисбаланс загрузки

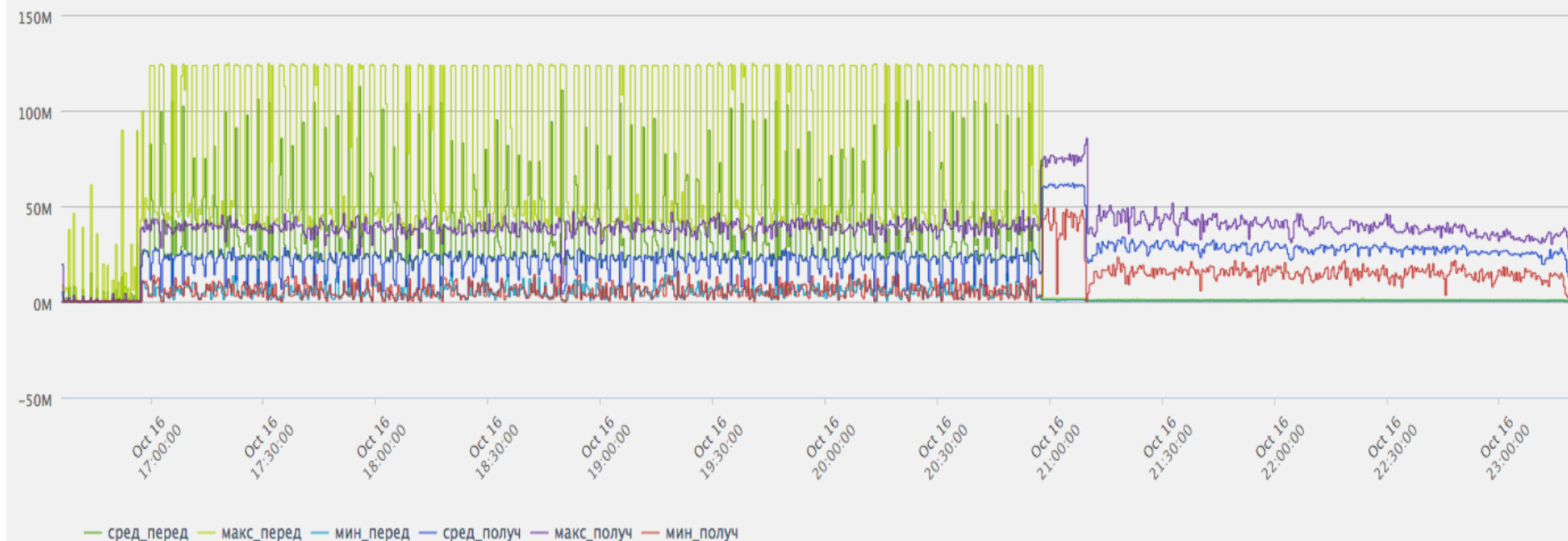
Анализ динамики выполнения суперкомпьютерных приложений

Ethernet (Panasas)

Скорость передачи данных по сети Ethernet (байт/сек)

analyze_eth_tpl8_hddmem-1381915736-24489.csv 00:03:11.397 (30.424 lines/sec) transform

Скорость передачи данных по сети Ethernet (байт/сек)



Интенсивный обмен через Ethernet, работа с ФС (запись)

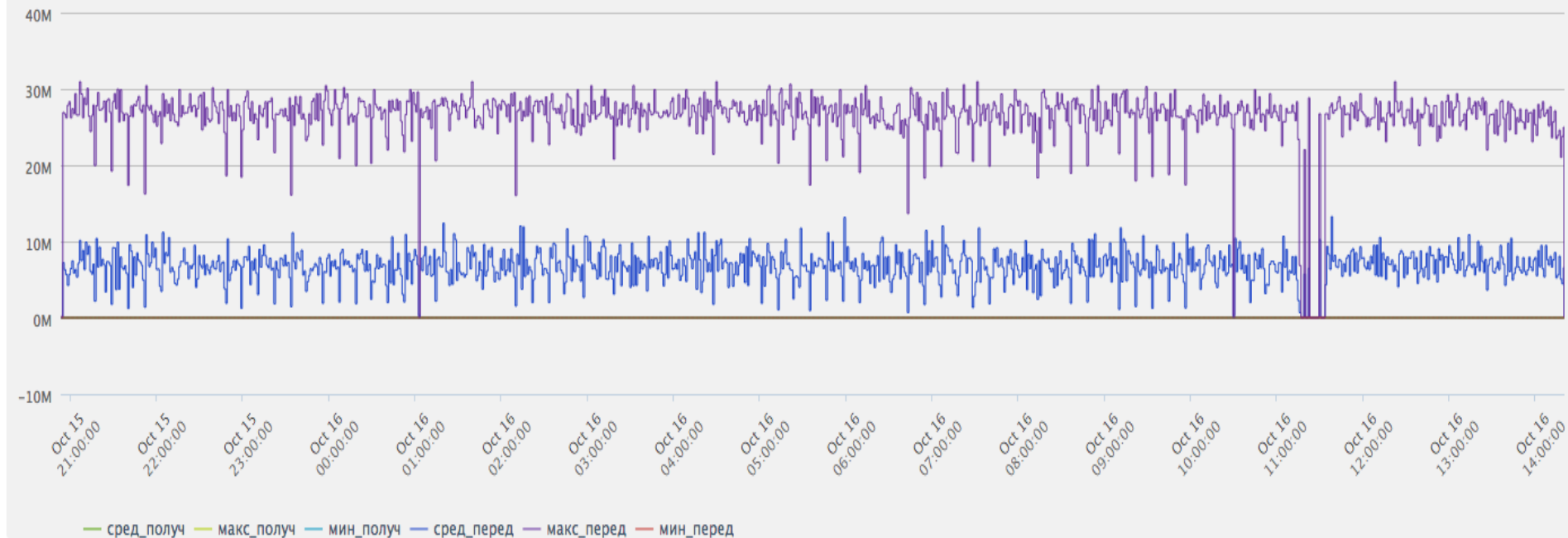
Анализ динамики выполнения суперкомпьютерных приложений

HDD I/O

Скорость обмена данными с диском (байт/сек)

analyze_disk_bytes_tpl8_hdd-1381870388-95140.csv 00:02:18.37 (43.401 lines/sec) transform

Скорость обмена данными с диском (байт/сек)



Очень высокая интенсивность записи, около максимальной для HDD

Общая схема метода анализа соответствия разделов суперкомпьютеров и суперкомпьютерных приложений

ПОДГОТОВИТЕЛЬНЫЙ ЭТАП

определение состава и получение интегральных характеристик по данным системного мониторинга



ПОИСК ПРИЗНАКОВ НЕЭФФЕКТИВНОГО ПОВЕДЕНИЯ ПРИЛОЖЕНИЯ В РАЗДЕЛЕ

- загрузка центрального процессора
- баланс загрузки вычислительных ядер
- использование профильного для раздела ресурса



ДЕТАЛЬНЫЙ АНАЛИЗ ПРИЛОЖЕНИЙ



**УТОЧНЕНИЕ
ПОЛИТИК ДОСТУПА**



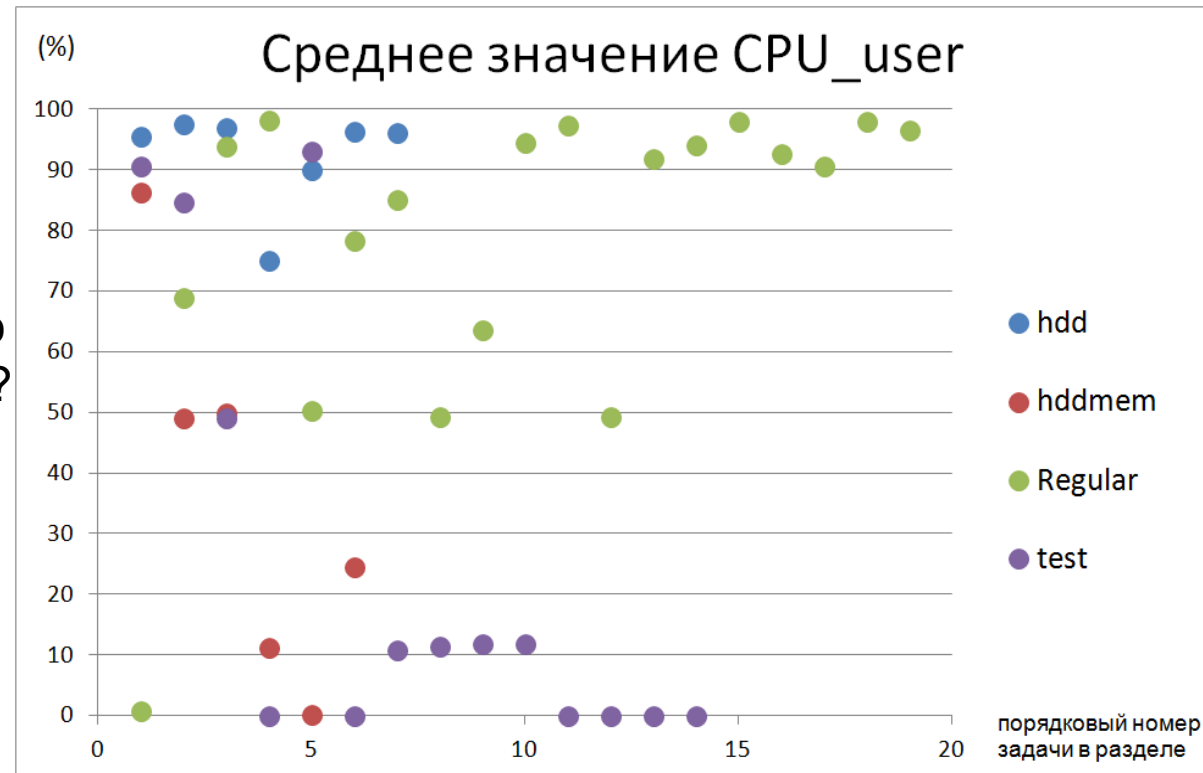
**КОРРЕКТИРОВКА
СИСТЕМНЫХ НАСТРОЕК**



**ОПТИМИЗАЦИЯ
ПРИЛОЖЕНИЙ**

Анализ соответствия суперкомпьютерных приложений разделам суперкомпьютера

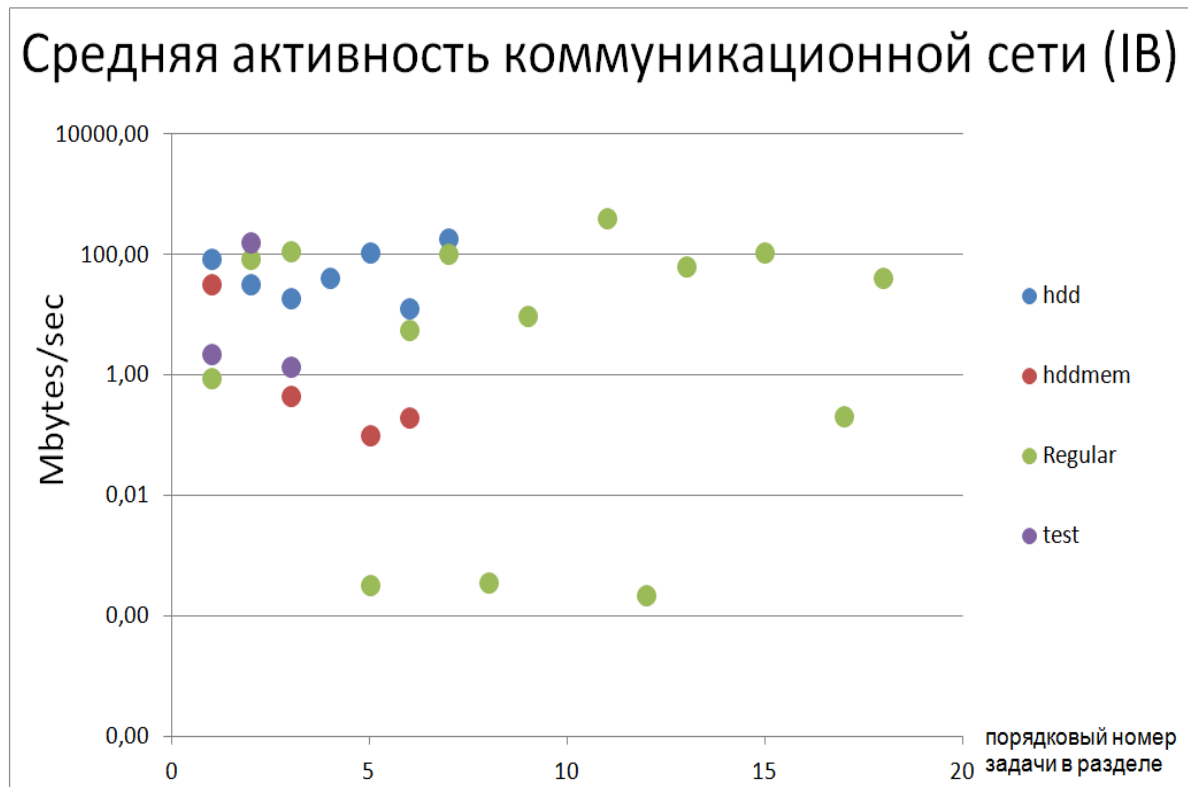
- Низкий уровень использования процессора в разделе TEST
- Три задачи в HDDMEM – может быть слишком много мелких операций с диском?
- Наиболее сбалансированные задачи – в разделе HDD
- Высокие в целом интегральные характеристики



Раздел	ядер	Память	диски
regular	4152	1 ГБ/ядро	нет
hdd	520	1 ГБ/ядро	есть
hddmem	256	2 ГБ/ядро	есть
test	80	1 ГБ/ядро	нет

Анализ соответствия суперкомпьютерных приложений разделам суперкомпьютера

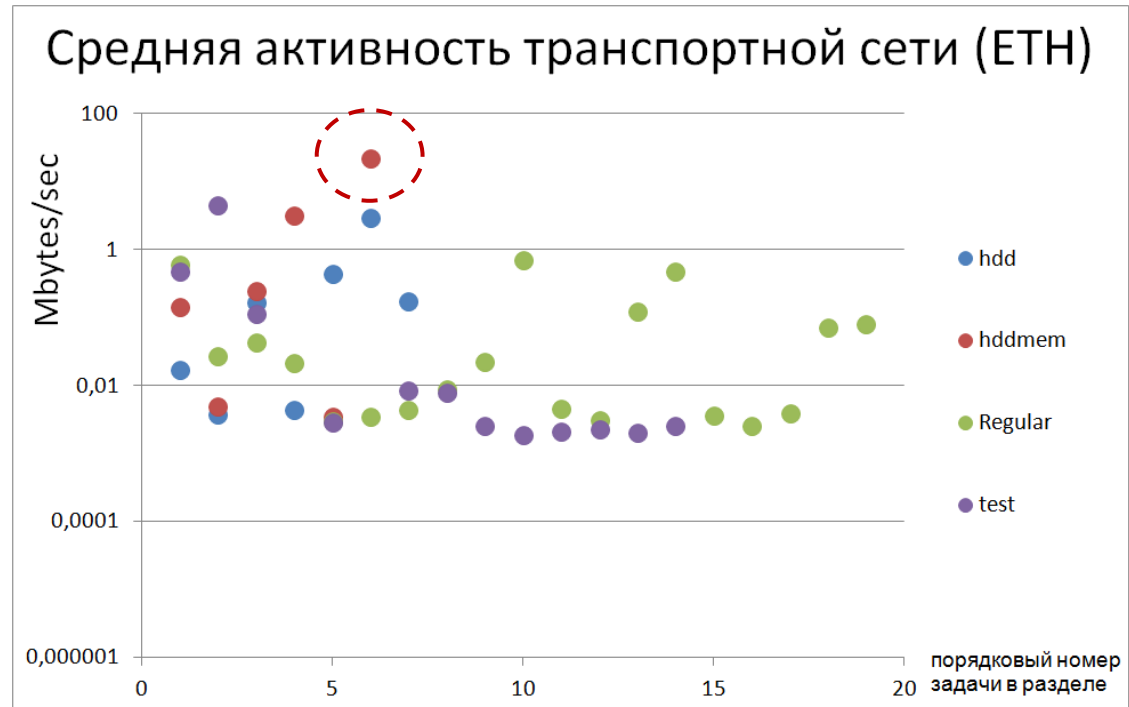
- В очереди TEST всего три задачи в принципе использовали Infiniband
- В очереди REGULAR три задачи почти не обменивались данными: нужен ли для их решения кластер вообще?
- Высокий уровень использования в целом



Раздел	ядер	Память	диски
regular	4152	1 ГБ/ядро	нет
hdd	520	1 ГБ/ядро	есть
hddmem	256	2 ГБ/ядро	есть
test	80	1 ГБ/ядро	нет

Анализ соответствия суперкомпьютерных приложений разделам суперкомпьютера

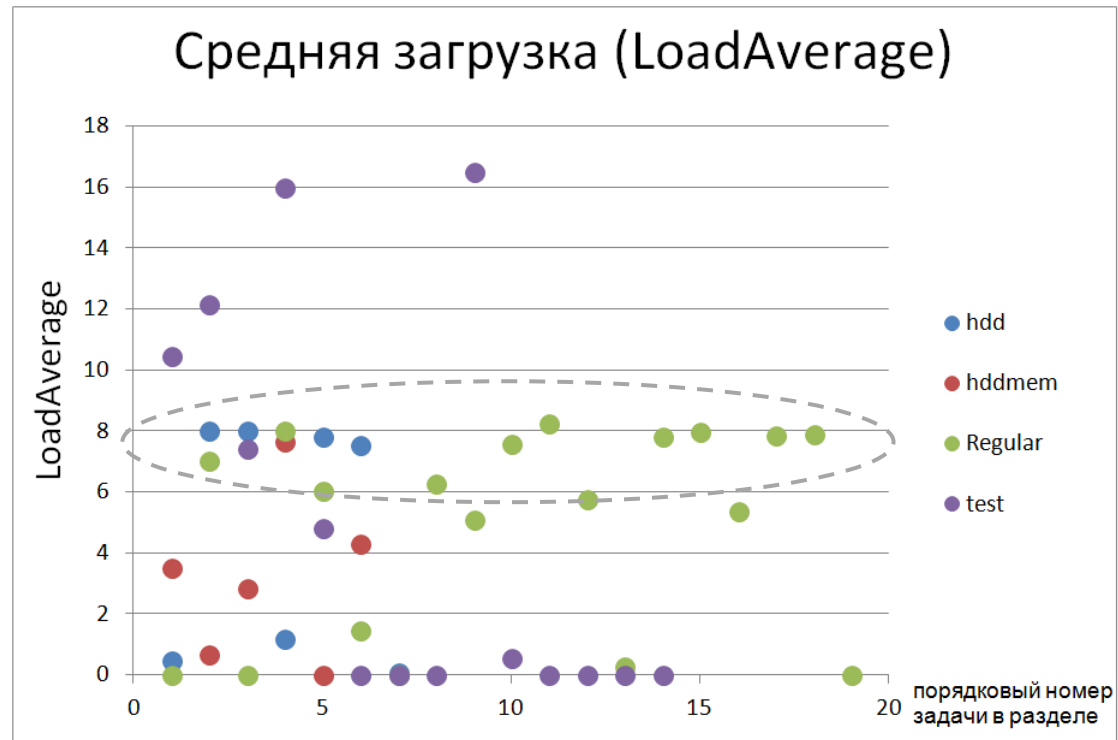
- Высокий уровень использования в HDDMEM может говорить о том, что задачи ставились в раздел не из-за оснащённости HDD, а из-за большей оснащённости оперативной памятью.



Раздел	ядер	Память	диски
regular	4152	1 ГБ/ядро	нет
hdd	520	1 ГБ/ядро	есть
hddmem	256	2 ГБ/ядро	есть
test	80	1 ГБ/ядро	нет

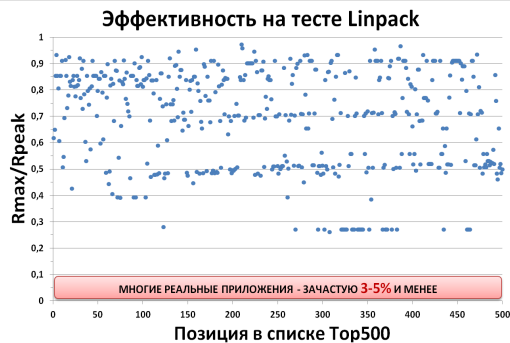
Анализ соответствия суперкомпьютерных приложений разделам суперкомпьютера

- Дисбаланс в TEST
- HDD и HDDMEM: только у одной задачи уровень загрузки высок, во всех остальных случаях есть подозрение на сильный дисбаланс загрузки
- Значения в районе 8 – хорошо для данного суперкомпьютера (на каждом узле установлено 2 процессора * 4 ядра)



Раздел	ядер	Память	диски
regular	4152	1 ГБ/ядро	нет
hdd	520	1 ГБ/ядро	есть
hddmem	256	2 ГБ/ядро	есть
test	80	1 ГБ/ядро	нет

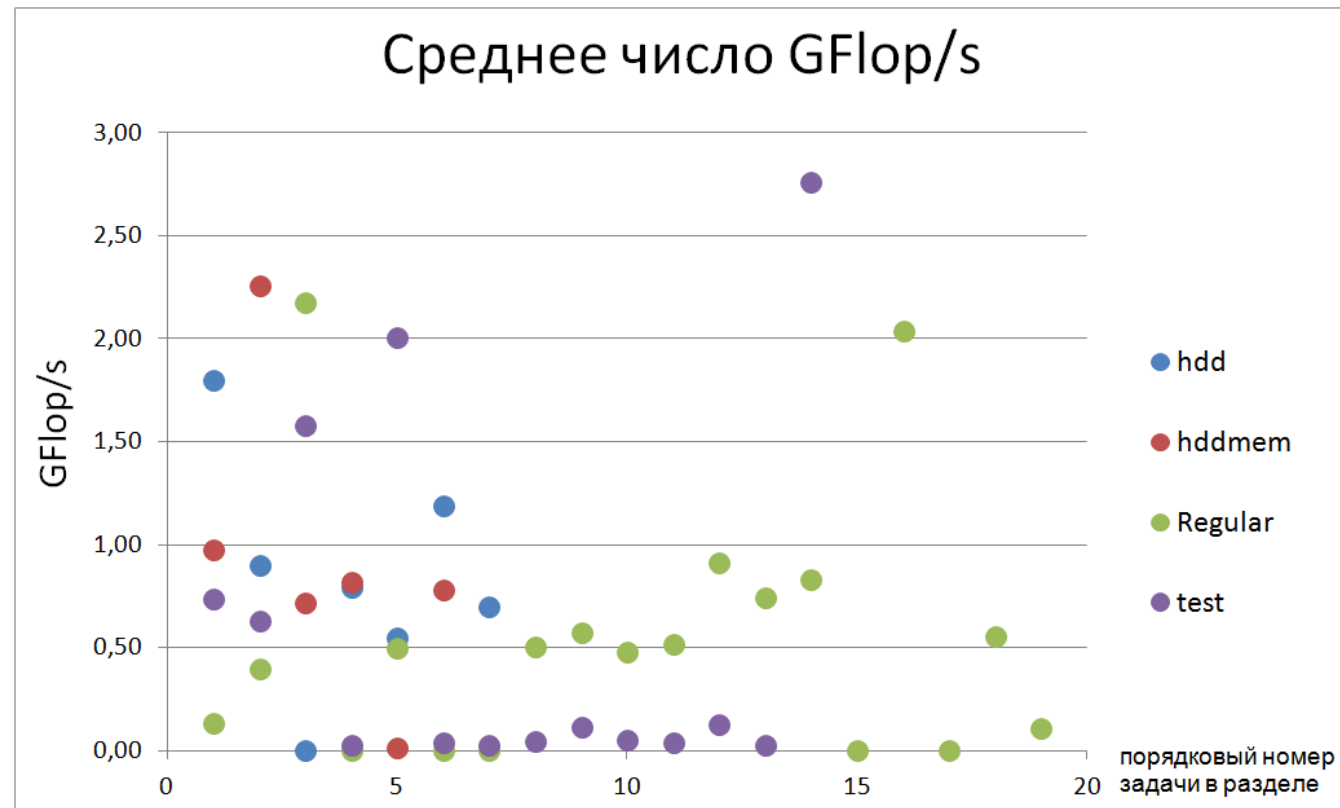
Анализ соответствия суперкомпьютерных приложений разделам суперкомпьютера



- Низкий уровень использования в разделе TEST

- В целом низкая интенсивность вещественной арифметики —

около 5% от пиковой
(12 GFlop/s)



Раздел	ядер	Память	диски
regular	4152	1 ГБ/ядро	нет
hdd	520	1 ГБ/ядро	есть
hddmem	256	2 ГБ/ядро	есть
test	80	1 ГБ/ядро	нет

Исследование характера использования ресурсов системы в целом

Два взаимодополняющих метода:

- На основе исследования интегральных характеристик задач
- На основе анализа усредненных данных системного мониторинга всех узлов (ядер, интерфейсов) счетного поля, без привязки к отдельным задачам

Исследование характера использования ресурсов системы в целом без привязки к отдельным задачам



Публикации по теме работы

В изданиях списка ВАК

1. Никитенко Д.А. Комплексный анализ производительности суперкомпьютерных систем, основанный на данных системного мониторинга // Вычислительные методы и программирование. 2014. 15. 85–97.
2. Антонов А.С., Жуматий С.А., Никитенко Д.А., Стефанов К.С., Теплов А.М., Шве́ц П.А. Исследование динамических характеристик потока задач суперкомпьютерной системы // Вычислительные методы и программирование. 2013. 14. 104–108.
3. Андреев Д.Ю., Антонов А.С., Воеводин Вад.В., Жуматий С.А., Никитенко Д.А., Стефанов К.С., Шве́ц П.А.. Система автоматизированного поиска ошибок и неэффективностей в параллельных программах. Вычислительные методы и программирование. 2013. 14. 48–53.
4. Антонов А.С., Воеводин Вад.В., Жуматий С.А., Никитенко Д.А., Стефанов К.С., Шве́ц П.А. Автоматизация поиска ошибок и неэффективностей в параллельных программах // Вычислительные методы и программирование. 2013. 14. 11–17.
5. Адинец А.В., Брызгалов П.А., Воеводин Вад.В., Жуматий С.А., Никитенко Д.А., Стефанов К.С. Jobdigest - подход к исследованию динамических свойств задач на суперкомпьютерных системах // Вестник Уфимского государственного авиационного технического университета, 2013. 17. №2. 131–137.
6. Адинец А.В., Брызгалов П.А., Воеводин Вад. В., Жуматий С.А., Никитенко Д.А., Стефанов К.С. Job Digest: an approach to dynamic analysis of job characteristics on supercomputers // Вычислительные методы и программирование. 2012. 13. 160–166.
7. Воеводин Вл.В., Жуматий С.А., Соболев С.И., Антонов А.С., Брызгалов П.А., Никитенко Д.А., Стефанов К.С., Воеводин Вад.В. Практика суперкомпьютера "Ломоносов" // Открытые системы. 2012. №7. 36–39.
8. Воеводин Вад.В., Стефанов К.С., Никитенко Д.А., Адинец А.В., Брызгалов П.А., Жуматий С.А. Hoplang - развитие языка обработки потоков данных мониторинга // Вычислительные методы и программирование. 2012. 13. 126–131.
9. Никитенко Д.А. Стефанов К.С. Исследование эффективности параллельных программ по данным мониторинга // Вычислительные методы и программирование. 2012. 13. 97–102.
10. Адинец А.В., Брызгалов П.А., Воеводин Вад.В., Жуматий С.А., Никитенко Д.А. Об одном подходе к мониторингу, анализу и визуализации потока заданий на кластерной системе. // Вычислительные методы и программирование, 2011. 12. 90–93.

В других изданиях

1. Mohr B., Voevodin V., Gimenez J., Hagersten E., Knuepfer A., Nikitenko D., Nilsson M., Servat H., Shah A., Winkler F., Wolf F., Zhujuv I. The HOPSA workflow and tools // Tools for High Performance Computing. Heidelberg: Springer, 2013. 127–146.
2. Адинец А.В., Брызгалов П.А., Жуматий С.А., Никитенко Д.А. Система визуализации параметров работы больших вычислительных систем // Тр. Междунар. конф. "Параллельные вычислительные технологии-2012 (ПаВТ-2012)". Новосибирск, 26–30 марта 2012 г. Челябинск: Издательский Центр ЮУрГУ, 2012. 714.
3. Адинец А.В., Жуматий С.А., Никитенко Д.А. Hoplang – язык обработки потоков данных мониторинга. // Тр. Междунар. конф. "Параллельные вычислительные технологии-2012 (ПаВТ-2012)". Новосибирск, 26–30 марта 2012 г. Челябинск: Издательский Центр ЮУрГУ, 2012. 351–359.
4. Жуматий С.А., Стефанов К.С., Антонов А.С., Жолудев Ю.А., Воеводин В.В., Никитенко Д.А., Адинец А.В. Подход к анализу эффективности больших вычислительных систем и программ // Тезисы докладов международной конференции "Современные проблемы математики, информатики и биоинформатики". Новосибирск, 11–14 октября 2011. 2011. 61.
5. Адинец А.В., Брызгалов П.А., Воеводин В.В., Жуматий С.А., Никитенко Д.А. Мониторинг, анализ и визуализация потока заданий на кластерной системе // Материалы XI Всероссийской конференции "Высокопроизводительные параллельные вычисления на кластерных системах". Нижний Новгород: Изд-во Нижегородского государственного университета, 2011. 10–14.

Основные результаты

1. Предложен и обоснован комплексный подход к анализу эффективности работы суперкомпьютерных приложений и систем на основе исследования потоков данных системного мониторинга суперкомпьютерных комплексов.
2. Разработаны методы анализа эффективности работы параллельных приложений и суперкомпьютерных систем, опирающиеся на исследование динамических и интегральных характеристик суперкомпьютерных приложений и данные системного мониторинга.
3. Предложена методика, опирающаяся на понятие причин и признаков снижения эффективности суперкомпьютерных приложений и систем, направленная на нахождение истинных источников потерь производительности.
4. Предложенные в данной работе подходы, методы и методики прошли успешную апробацию в рамках суперкомпьютерного комплекса Московского университета на большом числе реальных приложений пользователей.