С.А.Петунин (ФГУП «ВНИИА»)

Статистические шаблоны и метрики оценки рабочей нагрузки корпоративных суперкомпьютерных центров

Аннотация

В докладе предложены некоторые методические рекомендации для организации статистической обработки информации, получаемой из системы управления заданий вычислительного кластера в условиях проведения массовых расчетов. Предполагается, что кластер обслуживает локальных пользователей и имеется возможность получения дополнительной информации о характере приложений.

Ключевые слова и фразы: *Рабочая нагрузка, многопроцессорный вычислительный кластер, массовые расчеты, статистический анализ*

1 Введение

Технология обеспечения массовых расчетов на высокопроизводительных ЭВМ в нашей стране сформировалась еще в 70-80-х годах прошлого века. Тогда в крупных академических и оборонных научных центрах появились прообразы современных кластеров - «многомашинные» вычислительные комплексы (МВК). Эти комплексы требовали эффективного распределения ресурсов нескольких однородных вычислителей, в качестве которых обычно использовалась легендарная отечественная ЭВМ БЭСМ-6. Для МВК, как систем с общим вводом данных, общей дисковой памятью и единым управлением, было разработано специальное системное программное обеспечение (СПО) [1], включающее в себя, в частности, средства сбора статистики выполнения прикладных расчетов. Организационная модель поддержки вычислений МВК являлась корпоративной (в современной терминологии), т.е. ориентированной только на задачи собственных научных институтов, в противоположность концепции вычислительных центров коллективного пользования, которые обслуживали несколько организаций и которые появились вместе с развитием телефонных и локальных сетей доступа.

В настоящее время корпоративный режим эксплуатации современных многопроцессорных МВК продолжает оставаться основным, несмотря на расширяющуюся практику проведения удаленного моделирования на базе коммерческих или университетских суперкомпьютеров. Среди причин, обуславливающих автономность существования суперкомпьютерного центра, следует отметить потребности пользователей в регулярности и оперативности расчетов, их привязке к научно-производственному циклу предприятия и данным информационных баз, обеспечении защиты информации и т.д. Важная специфика функционирования такого центра заключается в уменьшении случайного фактора формирования рабочей нагрузки в сравнении с суперкомпьютерами общего доступа, что приводит к лучшему пониманию поведения динамических процессов, протекающих в системе. Как следствие, появляются дополнительные уникальные возможности для повышения эффективности использования ресурсов МВК.

2. Организационно-техническая схема проведения массовых расчетов

Рассмотрим следующие специальные условия эксплуатации:

1) имеется новый МВК с прогнозируемо-высоким показателем загрузки;
2) кластер обслуживает большую группу локальных пользователей;
3) СПО МВК состоит из типового набора свободно-распространяемых пакетов (например: Slurm, Ganglia, утилит тестирования оборудования и т.п.)

Следует отметить, что возможности встроенной в компоненты СПО штатной статистики (ШС), ориентированной в большей степени на задачи оперативного статистического анализа,

принципиально ограничены ее специализацией, а также отсутствием аналитического аппарата исследования данных. При этом эксплуатационному персоналу сегодня необходимо обеспечивать управление чрезвычайно сложным техническим объектом с учетом взаимовлияния различных по природе факторов, например, состояния инженерного оборудования и выполнения прикладной параллельной программы. В связи с ограниченной функциональностью ШС, разработка и использование аналитического инструментария – актуальная возможность улучшения процесса управления массовым счетом.

На рис.1 представлен конкретный пример включения аналитической компоненты в организационно-техническую модель эксплуатации МВК.

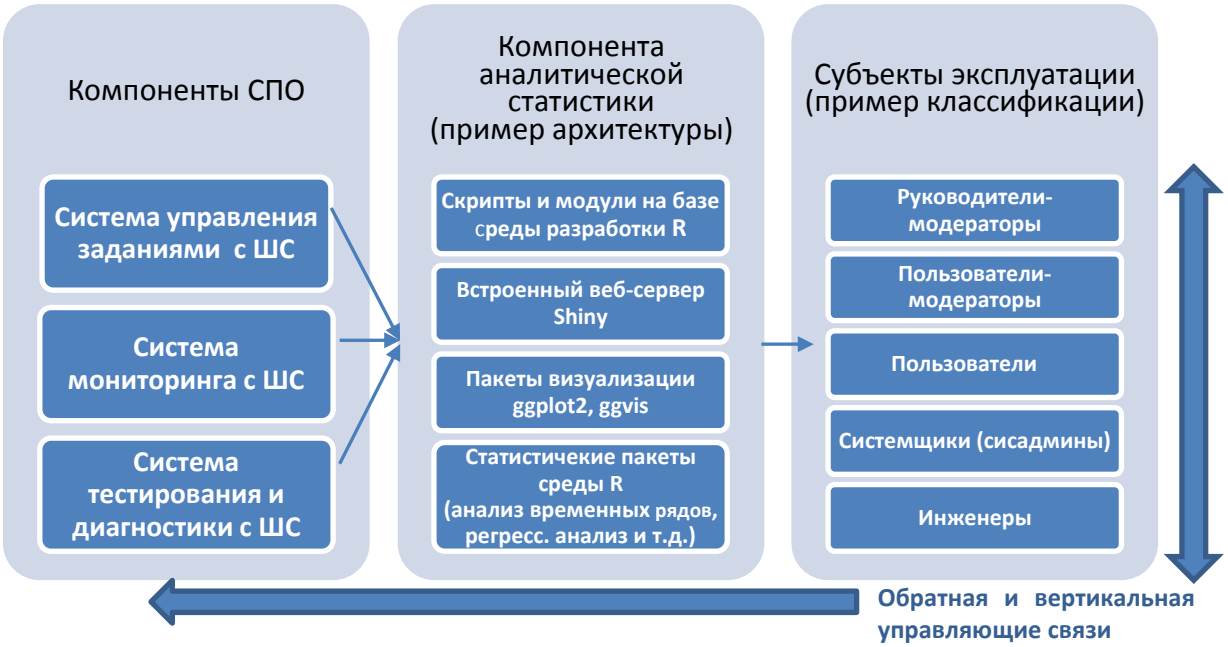

Рисунок 1. Место аналитической компоненты в схеме обеспечения эксплуатации МВК

Классификация групп сотрудников детализирована с учетом обеспечения поддержки массового счета и включает в дополнение к традиционному триадному расслоению (руководство, пользователи, технические службы) модерирующий персонал. К модераторам относятся как ответственные прикладные пользователи, управляющие группой расчетов, так и руководители-аналитики, которые счетом не занимаются, но принимают решения об изменении плана проведения расчетов. Исторически устареваемая группа «операторов» исключена из схемы, т.к. работа по снятию и запуску заданий либо осуществляется самими пользователями, либо (в условиях сложного контента очереди) требует квалифицированных действий, связанных с перестройкой вычислительной среды, которая обеспечивается при поддержке системных администраторов.

В дальнейшей части материала мы будет рассматривать только задачу проведения анализа рабочей нагрузки МВК. Здесь можно рекомендовать простейшую двухуровневую «оффлайн»-схему создания и представления разнообразных статистик о вычислительных заданиях:

- с помощью известной (но синтаксически неудобной) консольной команды Sacct, выбрать необходимые поля из базы данных планировщика Slurm, дополнив их информацией по событиям из файла slurmctld.log, и построить исходные таблицы для последующей обработки;

- использовать в качестве легкого инструмента для построения начальной статистической модели рабочей нагрузки, проведения постобработки и визуализации кроссплатформенную среду на базе языка программирования R, реализационная структура которой представлена выше на рис.1.

Таким образом, уже в начале эксплуатации суперкомпьютера можно не только получить информацию: кто, когда, где и сколько считал, но и обеспечить наглядной информацией связанный с эксплуатацией административный персонал. Создание многофункционального аналитического программного обеспечения, которое учитывало бы все типы нагрузок для различных информационных систем предприятия – большая и сложная задача, требующая привлечения математиков и ИТ-специалистов разного профиля. ФГУП «ВНИИА» начинает работы по этому проекту. Ниже приведены первые оценки возможностей среды программирования R и результаты анализа (пока без полноценного применения аппарата мат.статистики) некоторых выборок из рабочей нагрузки. Статистика, используемая в данной работе, взята из исторических баз менеджера ресурсов Slurm.

3 Характеристики рабочей нагрузки

3.1 Базовые временные характеристики многопроцессорных заданий

Как известно, единицей потребления вычислительных ресурсов МВК является «задание» (расчет). Совокупность поставленных на счет заданий создает рабочую нагрузку.

Время жизни любого задания *j* определяется тремя возрастающими пороговыми значениями: $T_{Submit}(j) < T_{Start}(j) < T_{End}(j)$, соответственно обозначающими времена формирования, запуска и завершения. На их основе вычисляются характеристики выполнения задания:

$T_{Wait} = T_{Start} - T_{Submit}$ (время ожидания)
$T_{Run} = T_{End} - T_{Start}$ (время выполнения, wall-time)
$T_{Response} = T_{Run} + T_{Wait}$ (время отклика).

Следует отметить, что геометрическое представление заданий как фиксированных прямоугольников, отображенных в пространство { CPU x Время }, в некоторых случаях может иллюстрировать только отдельный шаг задания, а не всю работу, помещенную пользователем с помощью командного файла в один из разделов-очередей системы управления заданиями, т.к. во время выполнения расчета возможно изменение объема выделяемых процессорных ресурсов (для пластичных или гибких заданий) [2]. Поскольку менеджеры ресурсов присваивают уникальные идентификаторы *j* каждому шагу, еще один важный временной показатель $T_{CPU}(j) = CPU(j)$ x $T_{Run}(j)$ определяет объем потребленных ресурсов. Значение $CPU(j)$, как правило, означает количество ядер, выделенных заданию, но существуют планировщики, оперирующие вычислительными единицами на уровне процессоров или узлов.

Принципиальными шаблонами рабочей нагрузки МВК являются поведение и календарные циклы входного потока заданий $\{T_{Submit}(j)\}$, распределение которого влияет на все процессы, протекающие в МВК. Из теории известно, что поток формирования заданий для суперкомпьютеров в большей степени имеет характеристики самоподобия, нежели соответствует классическому пуассоновскому процессу с постоянной интенсивностью событий [3]. Для разработки эффективных эвристических алгоритмов планирования параллельных заданий часто применяется моделирование рабочей нагрузки, которое использует распределение межинтервального промежутка прихода заданий $\{T_{Submit}(j+1) - T_{Submit}(j)\}$. Однако, для регулярно проводимого анализа с целью выработки политик загрузки МВК данный показатель не является информативным и не содержит специфические особенности для корпоративных или общедоступных вычислений.

На рис.2 приведена суммарная гистограмма распределения по часам суточного потока задний МВК, проводящего расчеты для отраслевого российского центра фундаментальных и прикладных исследований (ЦФПИ) [4]. Для сравнения добавлены относительно «свежие» исторические данные потока заданий международных суперкомпьютерных центров с аналогичной научно-исследовательской специализацией [5].

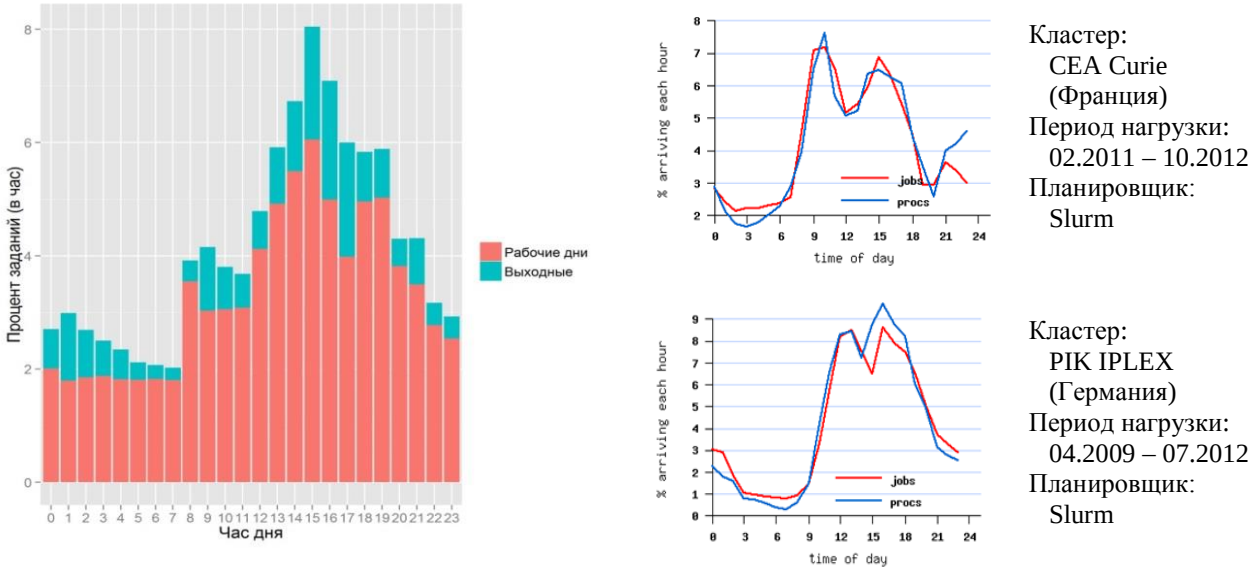

Рисунок 2. Патерны суточного потока заданий

Профили графиков формы «двугорбый верблюд» (справа) отражают интенсивность интерактивной работы пользователей, т.к. оба зарубежных научных центра могут обеспечивать дистанционную работу многих исследовательских рабочих групп. Характерной особенностью этих графиков является утренняя яма. У графика МВК (слева) утренний спад в рабочие дни несколько сглажен, в частности, по причине автоматического регулирования рабочей нагрузки скриптами пользователей и модераторов. Еще одним периодическим временным распределением является недельный цикл (рис.3).

Интенсивность входного потока следует анализировать в паре с утилизацией вычислительного поля. Данные показатели не всегда коррелируют друг с другом. При этом метрика утилизации МВК, которую общепринято считать статистическим показателем эффективности использования суперкомпьютерного оборудования, является в большей степени просто оперативным индикатором нераспределенных вычислительных узлов. Нераспределенность серверов может вызываться не только недозагрузкой, но и эксплуатационными событиями, например, исключением сегмента на время ремонта кондиционера. Для большинства промышленных суперкомпьютерных центров страны характерен высокий процент утилизации (более 90%).

Если детализировать утилизацию с точки зрения содержания расчетов, то стопроцентная загрузка всех процессоров не обязательно может обеспечиваться заданиями с полезными результатами. Кроме того, чрезвычайно важно выявление неравномерности утилизации процессорных ядер одним заданием. Этот анализ необходимо проводить автоматически в оперативном режиме в рамках мониторинговых или трассировочных сервисов МВК [7].

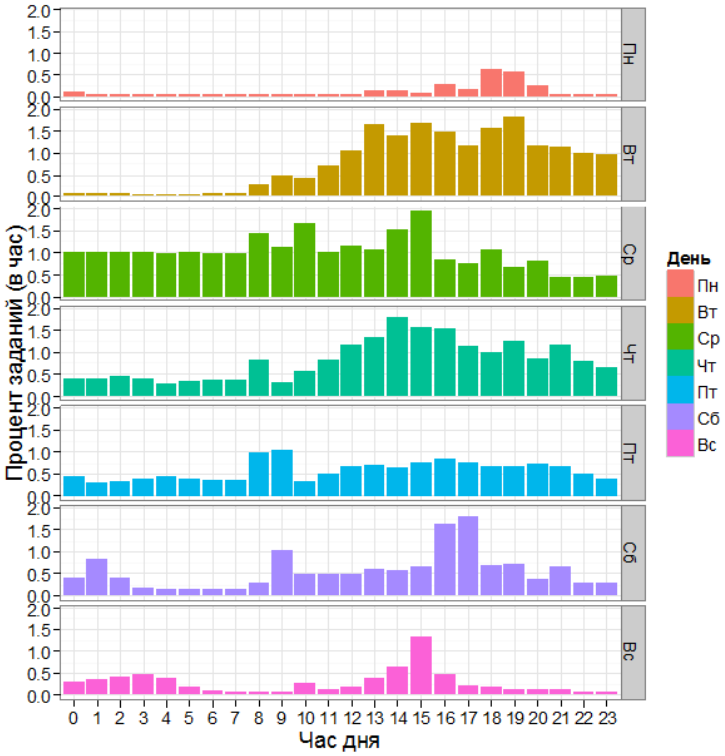

Рисунок 3. Распределение суточного потока по дням недели

3.2 Структурирование вычислительных ресурсов

В настоящее время существуют различные практики выбора количества разделов для планировщика заданий. Как правило, обязательно создается раздел, в рамках которого считаются «короткие» отладочные задания, лимитированные временной квотой (например, 10 минутами). Иногда по отдельным разделам структурируются «короткие», «средние» и «длинные» задания. Существуют примеры дифференциации разделов по размеру заданий, т.е. количеству запрашиваемых процессорных ресурсов, по типу вычислительных узлов и т.д. В качестве крайних политик по созданию разделов можно привести примеры из уже упоминаемого выше [5] архива суперкомпьютерных нагрузок. Суперкомпьютер Ливерморской национальной лаборатории LLNL Thunder имеет всего 2 раздела: «pbatch» и «pdebug», в то время как у приведенного на рис.2 CEA Curie для выполнения заданий создано 33 раздела.

В связи с гетерогенностью и гибридностью современных МВК целесообразно организовывать отдельные разделы для сегментов вычислительного поля, которые отличаются объемом оперативной памяти или наличием графических ускорителей.

3.3 Дополнительные показатели исторических данных

Системы управления заданиями сохраняют достаточно много параметров, характеризующих потребление расчетом различных ресурсов МВК: оперативной и дисковой памяти, коммуникационных сетей, процессоров и т.д. Однако, среди полей, идентифицирующих само приложение, чаще всего запоминаются только имена пользователя и файла со сценарием запуска. По таким данным сложно выстроить зависимость между поведением задания и типом расчета. Поэтому в крупных вычислительных центрах еще с до-суперкомпьютерных времен вводилась типизация расчета. Например, запись со статистикой

задания дополнялась иерархией: «*Проект*» –> «*Методика*» –> «*Задача*», которая не только улучшала анализ статистики, но использовалась для управления постановкой заданий на счет.

3.4 Маргинальное поведение

Одна из основных задач анализа любой статистической информации – выявление и устранение аномальных процессов в системе. Примером источника аномалий могут служить пользователи-«маргиналы», характеризующиеся, в частности:

- гиперактивностью;
- большим процентом ошибочного завершения задач;
- постоянным завышением запрашиваемого времени счета задания;
- низким средним T_{CPU} на задание;
- и т.д.

Для детального понимания подобных эксцессов рабочей нагрузки полезно анализировать кроме типовых штатных параметров log-файлов также специально вычисленные оценочные значения. В табл.1 приведены некоторые из них.

Показатели и статистики	Использование
Коэффициент замедления задания (slowdown) $K_{Sld}(j) = T_{Response}(j) / T_{Run}(j)$	Высокое значение характеризует большую долю времени нахождения задания в очереди. Группировка K_{Sld} для отдельного пользователя может выявлять необоснованность запросов на ресурсы или сигнализировать о заторах в очередях. Аномальность требует изменения политик планирования, регулирования входного потока или динамического расширения полосы ресурсов в разделе.
Коэффициент соответствия актуального и запрошенного времени исполнения задания $K_{Wall}(j) = T_{Run}(j) / T_{ReqWall}(j)$	Важный параметр для повышения эффективности работы планировщика (до 25-30%) [6]. Необоснованное завышение запрашиваемого времени приводит к падению производительности штатного алгоритма Backfill. Аномальность требует применения организационных механизмов «воспитания» пользователей или включения в планировщик автоматической коррекции на основе прогнозирования для данной методической задачи.
Среднее потребление CPU на запуск для пользователя $K_{CPU}(u) = mean(T_{CPU})$	Низкое значение для пользователя в сочетании с большим количеством запусков говорит о необоснованной сверхактивности. Аномальность требует установки квоты на одновременный запуск заданий одним пользователем.

Таблица 1. Пример использования оценочных коэффициентов

3.5 Распределение потребления ресурсов и представление статистических данных

Формы статистической информации, которые эксплуатационная служба обычно готовит для руководителей, чаще всего представляют собой простые выборки потребления конкретного ресурса с их визуализацией в виде линейных графиков, баров и круговых диаграмм. При этом содержание представления статистики легко расширить, если использовать современные инструменты, например, пакет ggplot2 статистической среды R. Построение функций распределения показателей, коррелограммы связности двух характеристик, тепловые карты и прочие визуализационные представления (рис.4-9) – все это помогает обнаружить устойчивости, закономерности или особые моменты поведения динамических процессов МВК, которые ведут к возникновению аномалий.

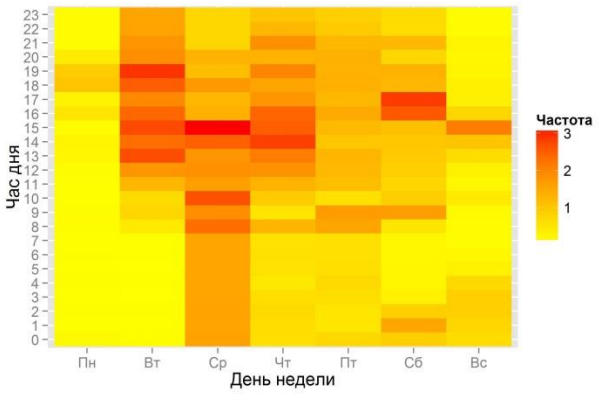

Рисунок 4. Тепловая карта частот распределения заданий в прямоугольнике {день недели, час дня}

Иллюстрирует сдвиг активности потока на вторую половину дня.

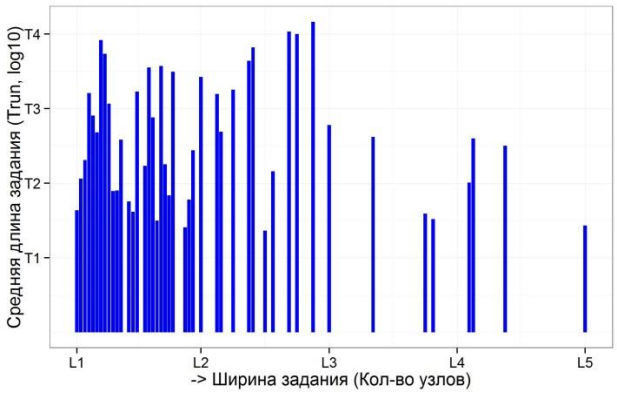

Рисунок 5. Бары среднего времени исполнения заданий на шкале кол-ва запрошенных узлов

Иллюстрирует отсутствие увеличения средней длины задания от его ширины.

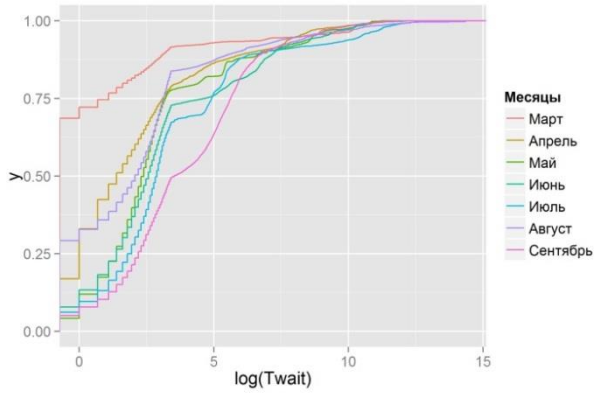

Рисунок 6. CDF функция распределения времени ожидания по месяцам

Иллюстрирует динамику нахождения заданий в очередях

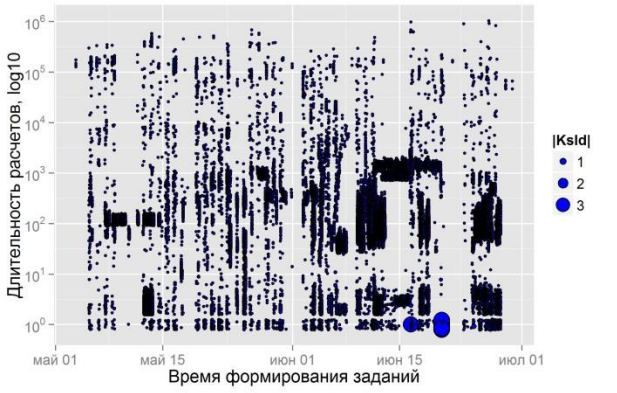

Рисунок 7. Точечный график распределения заданий в выбранном интервале времени

Иллюстрирует высокую интенсивность и затор во 2-й половине июня.

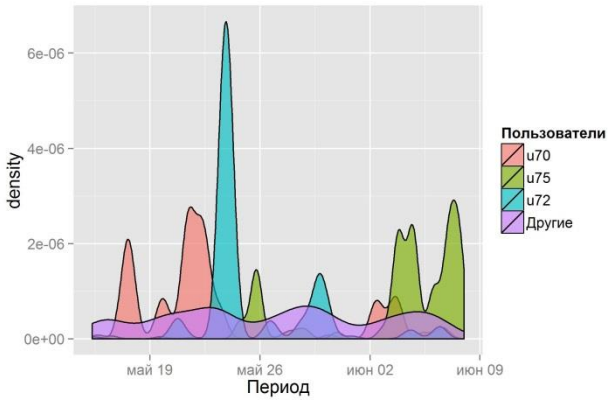

Рисунок 8. Плотность распределения заданий для группы пользователей в заданном диапазоне

Выявляет сверхактивных пользователей.

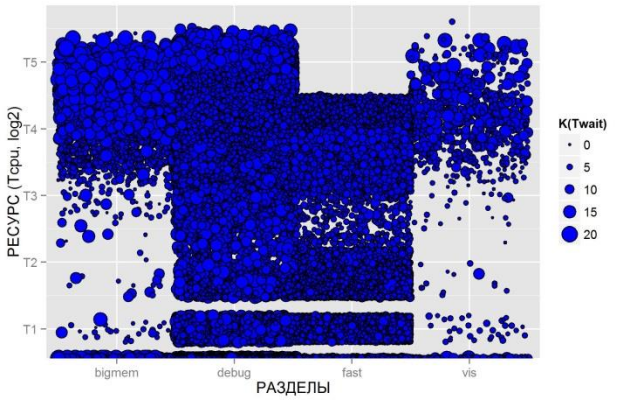

Рисунок 9. Коррелограмма распределения процессорного времени по разделам

Иллюстрирует высокие запросы на процессорный ресурс для раздела bigmem и связь с времен ожидания.

4 Заключение (проблема эффективного анализа)

Технологии получения статистической информации и ее использования традиционно связаны только с базой данных системы управления заданиями. Такие данные можно хорошо использовать для регулирования рабочей нагрузки. Однако, следует признать, что оптимальную загрузку ресурсов можно получить только в случае, когда ресурсы заказывает и распределяет не пользователь, а сама операционная система МВК, обладая точным знанием или оценочным прогнозом: сколько какому заданию необходимо выделить ЦПУ-часов, в какой раздел его поместить, как изменять геометрию задания в течение жизненного цикла, как могут повлиять на него текущее состояние параметров инженерных систем и т.д.

Инструментарий, применяемый сегодня при эксплуатации любых МВК, обязательно включает в себя средства мониторинга, диагностики оборудования, статистики и управления массовым счетом [7]. Независимость существования этих подсистем, вызванная различной природой объектов обслуживания, является серьезной эксплуатационной проблемой. Интеграция всех компонент системного программного обеспечения МВК в рамках общей информационной базы – это новая возможность для повышения эффективности работы суперкомпьютеров и производительности труда пользователей.

В качестве прогнозных выводов можно констатировать два тезиса:

1. Развитие средств анализа динамических процессов супер-ЭВМ, в частности вычислительной нагрузки, будет постоянно продолжаться, т.к. это один из базовых элементов построения современной технологии проведения массовых расчетов.
2. Можно предположить, что именно разработка интеллектуального управления МВК с интеграцией системных сервисов от уровня пользователя до аппаратуры позволит решить главную проблему эффективного распределения ресурсов: проблему стохастического характера вычислительной нагрузки.

Библиографический список

1. Мазурин Ю.Н., Охрименко Г.П., Петунин С.А., Функциональная структура операционной системы процессора управления данными. // ВАНТ, серия «Методики и программы численного решения задач мат.физики», 1983, вып., 3(4), с.59-61.
2. Новиков А.Б., Петунин С.А., Влияние специализированных алгоритмов планирования заданий на эффективность использования вычислительных ресурсов в частных случаях. // Труды XIII международного семинара «Супервычисления и математическое моделирование», РФЯЦ-ВНИИЭФ, 2011.
3. Feitelson, D.G., Workload modeling for performance evaluation // In Performance Evaluation of Complex Systems: Techniques and Tools, Springer-Verlag, 2002, pp. 114–141.
4. Сергеев О.В., Янилкин А.В., Метод анализа химического состава в молекулярно-динамических расчетах с потенциалом взаимодействия ReaxFF. // ВАНТ, серия «Математическое моделирование физических процессов». 2014. Вып. 3. С. 71-77.
5. Parallel workloads archive. URL http://www.cs.huji.ac.il/labs/parallel/workload/.
6. W.Tanga, N.Desaib, D.Buettnerb, Z.Lana, Job Scheduling with Adjusted Runtime Estimates on Production Supercomputers.
7. Petunin S.A., Ivanov K.V., Novikov A.B., Management of HPC Clusters: Development and Maintenance. // Proc. of the 15-th International Workshop on Computer Sciennce and Information Technologies CSIT’2013’, Vienna-Budapest-Bratislava, pp.43-46, 2013.