# ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ РАСПРЕДЕЛЕННЫХ РАСЧЕТОВ В СУПЕРКОМПЬЮТЕРНЫХ СРЕДАХ

***С.И. Соболев***

*Научно-исследовательский вычислительный центр МГУ имени М.В. Ломоносова*

Представлен подход к анализу эффективности расчетов в распределенных средах, построенных на основе суперкомпьютерных систем с использованием технологии X-Com. Подход может применяться при исследовании выполнения сверхбольших задач, для которых ресурсов одного суперкомпьютера оказывается недостаточно. Подход также позволяет получить статистическую информацию об особенностях прохождения заданий определенного класса в очередях суперкомпьютеров и выявить возможные проблемы распределения заданий по узлам вычислительных систем.

## Введение

В статье описываются результаты, полученные в рамках объединения двух направлений исследования методов организации распределенных расчетов с помощью системы метакомпьютинга X-Com [1, 2]. В 2010 г. система X-Com была адаптирована к работе совместно с основными системами управления потоками заданий суперкомпьютерных систем [3]. В 2011 г. было начато исследование и формирование системы оценок и характеристик распределенных неоднородных вычислительных сред с помощью инструментария на базе X-Com. Был рассмотрен простейший (базовый) случай организации двухуровневой распределенной среды (один сервер X-Com – множество клиентов X-Com, непосредственно работающих на вычислительных узлах) [4]. В данной работе приводятся результаты исследований свойств распределенных сред и расчетов в более сложном случае, когда задания от сервера X-Com поступают в системы очередей суперкомпьютеров через промежуточные серверы X-Com. Целью работы является обобщение и расширение методики оценки распределенных систем, полученной для простейшей архитектуры распределенных сред, на архитектуру, включающую промежуточные серверы X-Com и системы очередей, а также разработка инструментария для получения и анализа свойств распределенных систем.

## 1. Особенности многоуровневой архитектуры распределенной среды

Промежуточные серверы X-Com [2] работают с центральным сервером по тому же протоколу, что и клиенты X-Com, поэтому с точки зрения центрального сервера они ничем не отличаются от клиентов. Однако промежуточные серверы скрывают всю инфраструктуру, находящуюся под ними, будь то множество клиентов X-Com или система очередей. Если характеристики узла, на котором работает клиент X-Com – тип и частота процессора, объем оперативной памяти и т.д. – не меняются в течение расчета, то набор вычислительных ресурсов, координируемых промежуточным сервером, очевидно, может изменяться. Согласно протоколу взаимодействия клиентов и серверов X-Com, клиенты сообщают серверу характеристики вычислительных узлов, на которых они запущены, только при первом обращении. Параметры, передаваемые центральному серверу в качестве «характеристик узла», задаются в настройках промежуточного сервера перед началом его работы, они могут служить только в качестве примерного опи-

сания нижележащей вычислительной среды, так как не учитывают ее возможную динамику.

Еще одна проблема связана с наличием дополнительных накладных расходов при работе с системами очередей через промежуточные серверы. В базовой архитектуре основной источник задержек между отправкой порции данных клиенту на узел и получением результата их обработки – передача данных по сети (накладные расходы работы самого клиента достаточно малы). В многоуровневой архитектуре добавляется время ожидания порции во входном буфере промежуточного сервера, время ожидания задания в очереди и время ожидания готовой порции в выходном буфере (подробнее механизм буферизации описан в [2]). Эти параметры тесно связаны с настройками системы очередей и загруженностью вычислительной системы. Причем, если настройки системы очередей (максимальное число одновременно выполняющихся заданий и максимально допустимое число заданий в очереди) обычно прописаны в политиках доступа к вычислительному комплексу и известны заранее, то загрузка комплекса является, вообще говоря, динамическим фактором. В соответствии с политиками доступа настраиваются размеры буферов и другие параметры модуля взаимодействия промежуточного сервера X-Com с системой очередей на конкретной вычислительной системе.

Невозможным оказывается напрямую использовать показатели, связанные с производительностью (*MAXpeak, INTpeak, AVGpeak*), поскольку они опираются только на данные, полученные из настроек промежуточных серверов. Коммуникационная эффективность расчета *Ecomm* окажется довольно низкой, поскольку она учитывает неизбежные накладные расходы на ожидание заданий в очереди и порций данных во входных/выходных буферах. Из-за буферизации также снизится комплексная эффективность *Ecomplex*. Для того, чтобы разобраться в структуре такого расчета, протоколов только центрального сервера оказывается недостаточно. Необходим анализ событий всех промежуточных серверов, участвовавших в расчете. Для реализации этого функциональность промежуточных серверов X-Com была расширена возможностью протоколирования своих действий в лог-файлы. Для анализа таких файлов был разработан специальный инструментарий.

## 2. Вычислительный эксперимент

Проиллюстрируем исследование характеристик распределенного расчета на примере решения модельной задачи в распределенной среде, состоящей из 4-х суперкомпьютеров. Характеристики задействованных вычислительных ресурсов приведены в таблице 1.

Таблица 1. Список суперкомпьютеров, составлявших основу распределенной вычислительной среды

| **Суперкомпьютер** | **Вычислительные узлы** | **СУПЗ** | **Очередь** | **Число заданий в очереди** |
|---|---|---|---|---|
| «Ломоносов» (МГУ, Москва) | 2x Intel Xeon 5570 2.93 ГГц, 12 Гб | Slurm | regular4 | 20 |
| | | | test | 20 |
| «Чебышев» (МГУ, Москва) | 2x Intel Xeon E5472 3.0 ГГц, 8 Гб | Cleo | regular | 10 |
| | | | test | 10 |
| СКИФ Урал (ЮУрГУ, Челябинск) | 2x Intel Xeon E5472 3.0 ГГц, 8 Гб | Torque | batch (по умолчанию) | 20 |
| СКИФ Cyberia (ТГУ, Томск) | 2x Intel Xeon 5150 2.66 ГГц, 8 Гб | Torque | batch (по умолчанию) | 20 |

Параметры модельной задачи были подобраны так, чтобы время обработки одной порции составляло не более 10 минут (при общем числе порций 5000). Это позволило

использовать тестовые очереди суперкомпьютеров МГУ в дополнение к основным очередям. На головных машинах этих суперкомпьютеров было запущено по два промежуточных сервера X-Com, каждый из которых работал со своей очередью. На суперкомпьютерах СКИФ Урал и СКИФ Cyberia использовались только очереди по умолчанию, на их головных машинах работало по одному промежуточному серверу X-Com. Центральный сервер X-Com располагался на головной машине суперкомпьютера «Чебышев».

Рассмотрим характеристики [4], полученные в результате анализа протокола работы центрального сервера. Расчет продолжался 4 часа 39 минут. Ускорение *Speedup* достигло значения 70.5 – это можно считать хорошим результатом с учетом того, что в рамках задачи одновременно работало не более 100 процессов. Коэффициент ускорения *Cspeedup* составил 11.75, но в данном случае это значение нельзя считать корректным, поскольку с точки зрения центрального сервера над расчетом работало всего 6 процессов. Комплексная эффективность *Ecomplex* составила 95.92% – снова хороший результат: размеры входных и выходных буферов промежуточных серверов были малы по сравнению с общим числом порций в задаче, поэтому избыточные обмены и перерасчеты оказались незначительными на общем фоне. А вот коммуникационная эффективность *Ecomm*, как и предполагалось, оказалась достаточно низкой – 35.07%. Эта величина отражает временны**ы**е затраты на организацию всего расчета. Чтобы разобраться в структуре этих накладных расходов и причинах низкого уровня этого показателя, необходим анализ потоков событий каждого из промежуточных серверов. Для этого построим графики на основе табличных файлов, генерируемых анализаторами лог-файлов серверов X-Com.

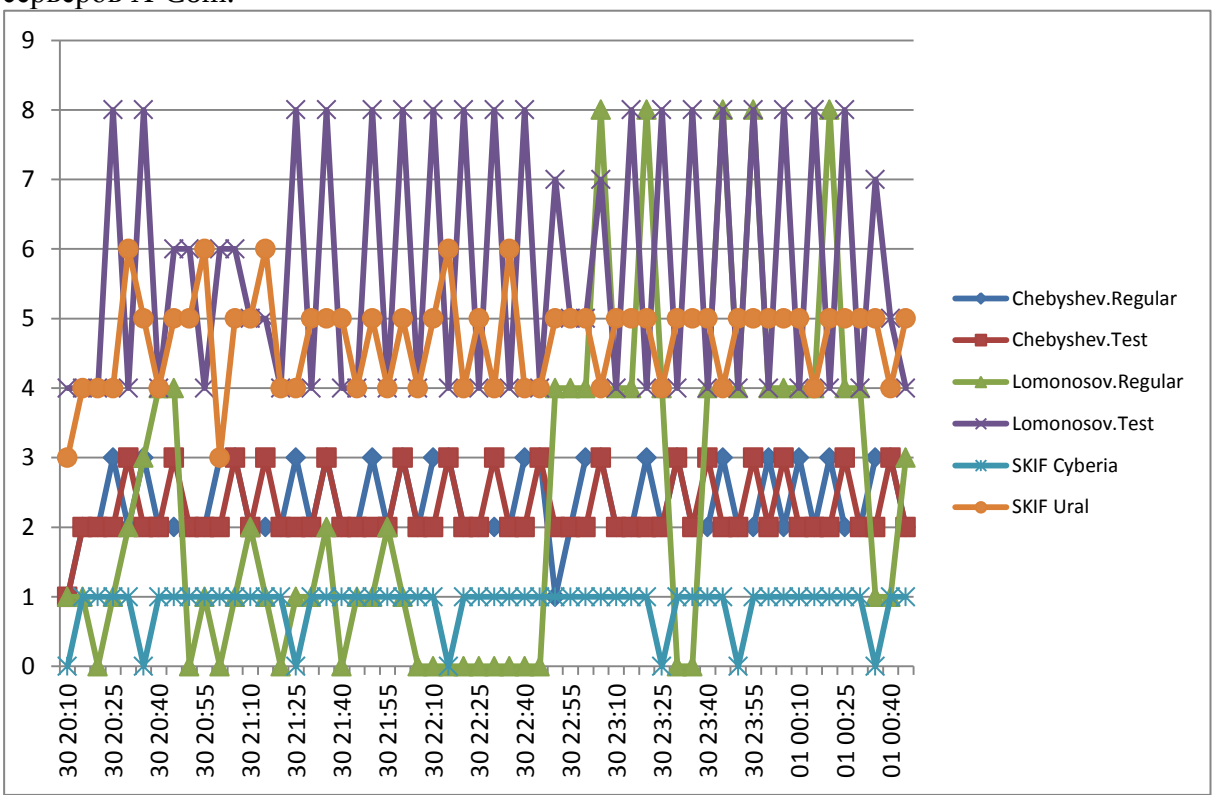


Рис. 1. График интенсивности отправки результатов центральному серверу

Рисунок 1 демонстрирует интенсивность получения центральным сервером результатов обработки порций данных от промежуточных серверов. По вертикальной оси отложено число запросов типа ASW, в каждом из которых передавалось по 5 результирующих порций.

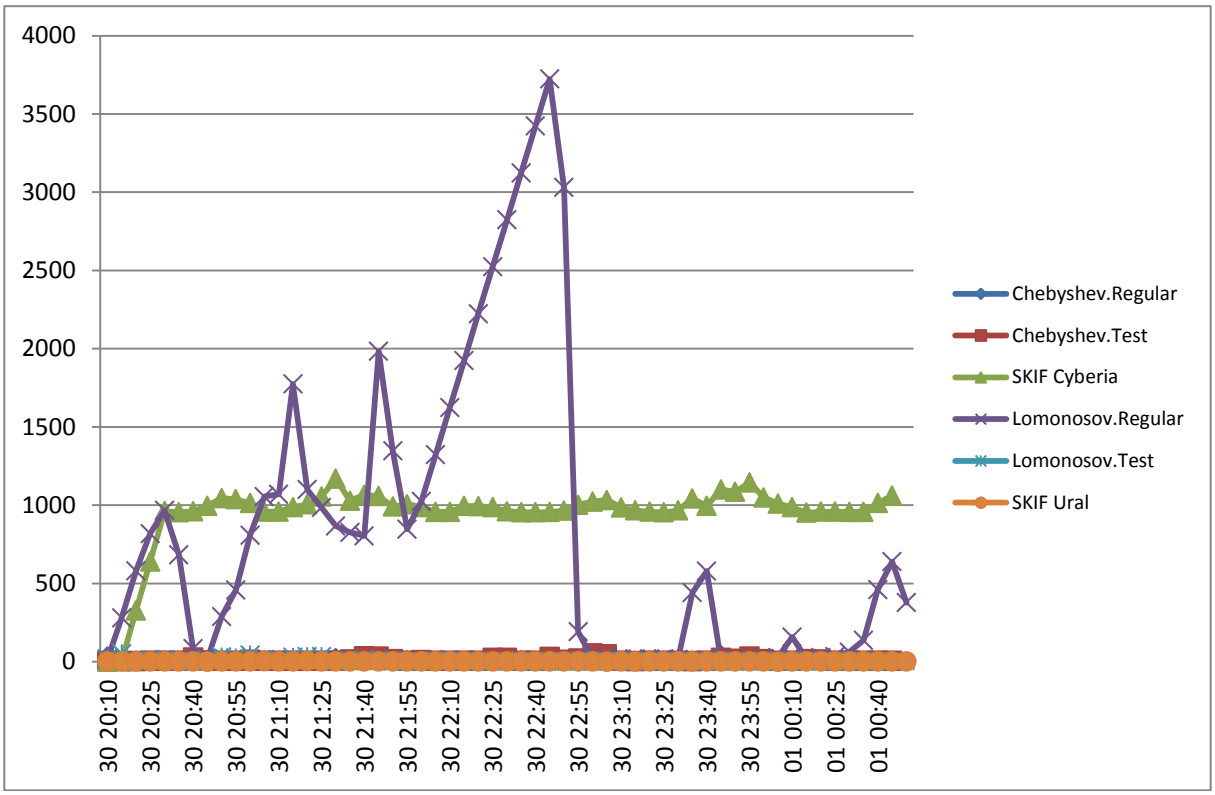


Рис. 2. Среднее время ожидания заданий в очереди

Среднее время ожидания заданий в очередях суперкомпьютеров приведено на рис. 2 (значения по вертикальной оси отложены в секундах). Хорошо видна постоянная и однородная загрузка очереди суперкомпьютера СКИФ Cyberia, а также интенсивное использование основной очереди «Ломоносова». Относительно небольшая загрузка была зафиксирована у тестовых очередей суперкомпьютеров МГУ, основной очереди «Чебышева», а также у суперкомпьютера СКИФ Урал. Эти данные хорошо коррелируют с оценками комплексной эффективности *Ecomplex*, полученными отдельно для каждой очереди: чем больше среднее время ожидания, тем ниже эти значения. Для «свободных» очередей оно составляло 44-48%, для «занятых» – 12-19%. Отметим, что понятия «загрузка», «свободно» и «занято» мы используем здесь неформально; их нельзя трактовать в широком смысле и обобщать на другие временн***ы***е интервалы, однако при описании ситуации в очередях на заданный момент времени для задач того же класса, что и модельная, они имеют вполне адекватную интерпретацию.

Интересный эффект был выявлен при анализе времен выполнения заданий. Поскольку с вычислительной точки зрения все задания были одинаковые, предполагалось получить одинаковые значения для каждого суперкомпьютера. Но это оказалось верно только для «Ломоносова». На остальных суперкомпьютерах время для большей части заданий было одинаковым и минимальным, времена же остальных укладывались в кластеры, которые и видны на графике (рис. 3) выше основных линий (по горизонтальной оси графика отложен номер порции, по вертикальной – время счета задания в секундах). Вероятно, такие задания попадали на узлы одновременно с другими заданиями. Такая ситуация может быть штатной (политики очереди могут допускать группировку нескольких однопроцессорных заданий на одном узле) или же ошибочной, свидетельствующей о некорректной работе механизма распределения заданий. На графике также заметно некоторое количество заданий с нулевым временем счета. Эти задания были завершены сразу после постановки на счет; причина такого поведения выясняется.

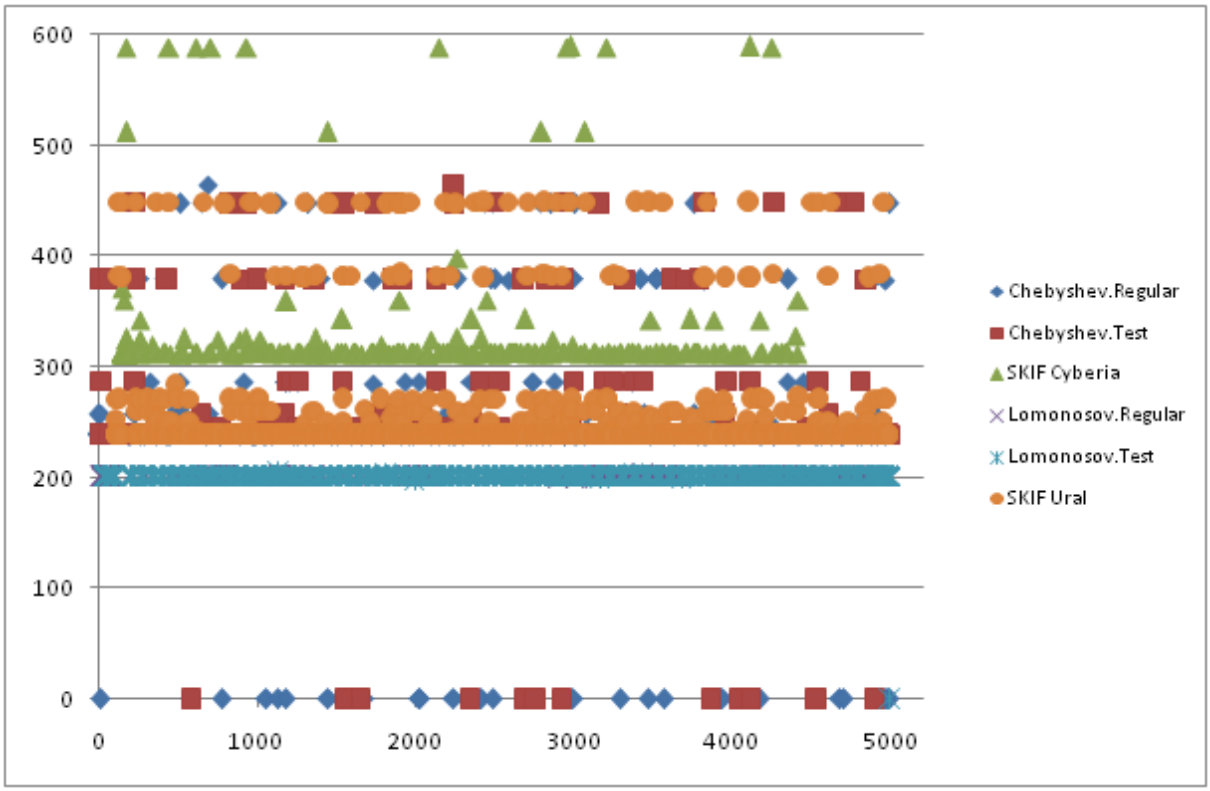


Рис.3. Время выполнения заданий на суперкомпьютерах

Таким образом, при анализе распределенных расчетов в многоуровневых средах следует использовать подход, отличный от анализа расчетов, выполняемых в рамках базовой архитектуры. В силу динамичности процесса многие характеристики также являются динамичными и могут быть представлены в виде графиков. Источником для них служат данные, формируемые инструментарием анализа потока событий серверов X-Com.

Работа выполняется при поддержке гранта Президента Российской Федерации для молодых российских ученых-кандидатов наук МК-5104.2011.9.

**Литература**

1. Система метакомпьютинга X-Com (официальный сайт) – [http://x-com.parallel.ru].
2. Воеводин Вл.В., Жолудев Ю.А., Соболев С.И., Стефанов К.С. Эволюция системы метакомпьютинга X-Com // Вестник Нижегородского государственного университета им. Н.И. Лобачевского. №4. 2009. С. 157–164.
3. Соболев С.И. Интеграция системы метакомпьютинга X-Com с системами управления прохождением заданий суперкомпьютерных комплексов. Распределенные вычисления и Грид-технологии в науке и образовании: Труды четвертой международной конференции (Дубна, 28 июня – 3 июля 2010 г.). С. 417–422.
4. Жолудев Ю.А., Соболев С.И., Стефанов К.С. Оценки и характеристики распределенных вычислительных сред. Высокопроизводительные параллельные вычисления на кластерных системах: Материалы XI Всероссийской конференции (Нижний Новгород, 1–3 ноября 2011 г.), С. 129–133.