Анализ основных тенденций в области хранения данных

Авторы: Тютляева Е.О., Московский А. А.

Аннотация: В статье анализируется тенденция возникновения и развития все большего числа научно-исследовательских проектов, накапливающих и анализирующих архивы данных большого объема (0,1-100 Пбайт). Проводится анализ возможностей современных суперЭВМ в части подсистемы хранения данных, на основе примерных характеристик вычислительных машин из верхней части рейтинга Топ 500. Проведен краткий анализ современных проектов развития средств хранения с перспективой применения на машинах эксафлопсного класса.

Ключевые слова: Системы хранения данных; приложения обрабатывающие большие объемы данных; тенденции развития.

Abstract: Examples of the most recent data-intensive applications are discussed. The recent advances in storage systems of the Top 500 supercomputers list are reviewed A brief review of recent progress in data storage tools with prospect to exascale systems is also given.

Keywords: Storage systems, Data-Intensive Computing, supercomputing.

Введение

Благодаря ускоренному развитию микроэлектроники, все возрастающее количество наблюдательных приборов все с высоким разрешением позволяет получать большие объемы данных (достигающие сотен терабайт и петабайт) в самых различных сферах человеческой деятельности, включая естественные науки, социологию и экономику. Современные мощности хранения позволяют сохранять и архивировать данные, которые могут представлять интерес для последующего исследования.

Под данными подразумеваются самые различные необработанные информационные материалы, включая не только снимки дистанционного зондирования, персональные медицинские данные, полные сырые данные различных наблюдений и экспериментов, но и базы данных социальных сетей, различных магазинов и прочую статистическую информацию. Кроме того, с ростом технологий возрастают требования к разрешению данных и новые вычислительные эксперименты предполагают более широкие временные и пространственные диапазоны обрабатываемых сырых данных.

Как утверждается в отчете "Большие данные: следующий рубеж для инноваций, соревнования и производительности" [1]: Данные становятся важнейшим фактором продукции сегодня - наравне с материальными активами и человеческим капиталом. В обозримом будущем экспоненциальный рост объема данных должен продолжиться в связи с возрастающей интенсивностью представления данных и сбора информации. Параллельно будет развиваться комплексное представление информации, объем социальных коммуникаций, количество

информации представленной в интернете. Большие объемы данных имеют значительные потенциал для того, чтобы стать значительной ценностью для бизнеса и пользователей.

По утверждениям IBM, каждый день создается около 15 PB новых данных ([2]). Это и научные данные, и сведения о проведенных операциях-транзакциях, и новые фотографии и отчеты в социальных сетях. Согласно информации из газет, для создания фильма "Аватар", потребовалась система хранения данных (далее в тексте СХД) более чем на 1 Петабайт ([3]). В социальную сеть Facebook каждый день добавляется около 12 TB данных (после сжатия) ([4]).

В книге "The Fourth Paradigm: Data-Intensive Scientific Discovery" [5], - исследование огромных массивов данных называют четвертой парадигмой науки, после экспериментальной, теоретической и вычислительной парадигм, сменявших друг друга на разных стадиях развития науки. Автором высказано предположение, что выявление закономерностей в больших массивах данных становится основным инструментом для исследования и получения новых знаний в передовых областях науки в наше время. К примеру, в [6] упоминается, что с 2001 до 2009 количество баз данных, зарегистрированных в Nucleic Acids Research увеличилось с 218 до 1,170.

В то же время, растущие объемы социальных данных способствуют расширению числа и масштабов исследовательских задач в области менеджмента, исследования рынка и социальной активности в виде аналогичных задач анализа взаимосвязей и закономерностей.

С увеличением интенсивности работы с данными во всех областях человеческой деятельности, характеристики подсистемы ввода-вывода и управления данными становятся одними из наиболее проблемных компонент современных информационных и вычислительных систем.

В данной статье мы постараемся провести краткий обзор ряда приложений, требующих интенсивной работы с данными, существующих решений в области высокопроизводительного хранения данных; оценить тенденции развития СХД, характеристик подсистем хранения данных наиболее мощных суперЭВМ и их соответствие реальным требованиям актуальных приложений.

Приложения, требующие интенсивной обработки больших объемов данных

В настоящее время проблема приложений, связанных с интенсивной работой с большими объемами данных находится на переднем крае науки. В англоязычной литературе такие задачи получили название “data intensive”, что можно перевести как «оперирующий большими объемами данных» (далее в тексте ОБОД). ОБОД называют те вычислительные задачи, в которых хранение, обработка и анализ значительных объемов данных становится первостепенной проблемой [7].

Сложность при обработке больших объемов данных порождает технологические проблемы как на уровне подсистемы хранения (скорость чтения/записи, надежность, доступных объем), так и на уровне обработки (доступные полосы пропускания оперативной памяти, возможный темп запросов в ОЗУ).

Для оценки готовности системы к обработке значительных объемов данных в 2010 году был анонсирован новый рейтинг, graph500 [8], который является первой серьезной попыткой дополнить список ТОП-500 оценкой возможностей системы для работы с большими объемными данными. Текущие тесты производительности, которые используются для построения рейтинга ТОП-500, не позволяют оценить пригодность высокопроизводительной установки для ОБОД приложений. Несмотря на то, что в ранжировании в новом списке пока приняли участие только 29 установок, уже очевидно, что результаты данного исследования значительно отличаются от рейтинга TOP-500. К примеру, лидирующий в GRAPH-500 суперкомпьютер Intrepid занимает всего 15 место в ТОП-500, и обгоняет по скорости работы с данными Jaguar (19 место в Graph-500 против 3 места TOP-500), Hopper (4 Graph-500 против 8 TOP-500), Jugene (2 Graph-500 против 12 TOP-500) и Lomonosov (3 Graph-500 против 13 TOP-500). Данный рейтинг позволяет оценить, прежде всего, готовность оперативной памяти системы к обработке приложений, интенсивно работающих с большими объемами данных. Тем не менее, даже этот рейтинг не дает возможности оценить системы хранения, характеристики которых играют ключевую роль для современных ОБОД приложений. Интересным курьезом является вхождение в текущую редакцию рейтинга машины из 1 узла суперЭВМ Kraken, использовавшего для хранения данных задачи, обычно размещаемым в ОЗУ, высокоскоростную СХД на твердотельных накопителях Fusion IO [55].

На уровне развития СХД перед исследователями стоят принципиально новые задачи. Наиболее значительным проектом, ставящим высокую планку для современных высокопроизводительных СХД, стоит назвать Большой Адронный Коллайдер в CERN. Миллионы сенсоров БАК генерируют около петабайта данных в секунду. В связи с тем, что современные мощности хранения не способны поддержать такие объемы хранения данных, большая часть измерений отфильтровывается на основе простейших правил. «Цель – не потерять ничего интересного». Тем не менее, даже после фильтрации и предобработки данных, коллайдер производит до 25 петабайт данных в год. Всего в центре обработки и хранения данных CERN 34 петабайта магнитных носителей и 45.3 петабайт на дисковых носителях. ([9])

Астрономия, наука непосредственно связанная с обработкой больших данных, также демонстрирует наличие актуальных проектов, связанных с обработкой значительных объемов данных. Приборы для астрономических наблюдений позволяют получать данные с все более высоким разрешением, исследователи заинтересованы в долгосрочном хранении полученных архивов для возможности последующих исследований. Кроме того, астрономические данные в большинстве своем не имеют ограничений приватности или коммерческой тайны, научное сообщество заинтересовано в общедоступности полученных данных, новых задачах, исследованиях и экспериментах, что накладывает дополнительные требования к распределенности и доступности данных.

Одним из примеров может быть система телескопов панорамного обзора и быстрого реагирования **Pan-STARRS,** нацеленная на обнаружение и изучение приближающихся к Земле

объектов, включая астероиды и кометы, которые могут оказаться опасными для нашей планеты. Одной из особенностей проекта является новаторская цифровая камера, позволяющая получать изображения 38,000 на 38,000 пикселей [10]. Каждое изображение, сделанное одной Pan_STARRS камерой, содержит около 2 Гбайт данных. В режиме полного обзора объем необработанных данных телескопа за ночь достигает нескольких терабайт.

В докладе **"Вычислительные вызовы в астрономии волн тяготения"** [11] упоминается, что в проекте участвует 4 детектора, расположенные на двух континентах, которые собирают необходимые данные для дальнейшего анализа и моделирования. Каждый детектор обладает скоростью передачи данных 10 MB/s. В докладе сказано, что годовой объем данных для 3-х детекторов составляет 947 TB. Основные проблемы, которые стоят перед исследователями, заключаются в управлении научными данными (в частности, полученными в результате анализа), лучшее управление научными потоками и недостаток квалифицированных кадров для разработки требуемой инфраструктуры.

Другим примером проекта, интенсивно работающего с данными, является NEEShub: Киберинфраструктура данных для моделирования землетрясений [12]. Целью данного проекта является создание национальной, многопользовательской исследовательской инфраструктуры для поддержки исследований и инноваций по минимизации ущерба от землетрясений и цунами. Под управлением проекта находятся значительные объемы разнообразных научных данных, включая изображения, видео, текст и т.п. По состоянию на март 2011 года в реализованной киберинфраструктуре находилось 417 проектов и почти 1 миллион файлов, объем данных менее 1 PB.

Проект, посвященный изучению нейронных связей, в настоящее время позволил получить 10 TByte данных – результатов моделирования, представляющих нейронные связи для примерно 1/80000 мозга мыши. Данный проект имеет значительный потенциал для масштабирования. Следующей целью является моделирование кубического миллиметра, что составит 1/1000 мозга мыши, и займет предположительно более 1 PB. Моделирование целой мыши по предположениям исследователей потребует системы хранения объемом в эксабайт. [13]

Большие объемы данных накапливаются и в климатологии. Например, немецкий центр по изучению климата (DKRZ, г. Гамбург) оснащен не только мощными суперЭВМ (более 150 ТФлопс), но и средствами визуализации данных, такими как специализированные комнаты, а также многоуровневой системой хранения данных, общим объемом около 60 Пбайт[54].

Приведенные примеры свидетельствуют о том, что задачи, находящиеся у переднего края науки, работают с большими объемами данных и уже сейчас имеют очень высокие и обоснованные требования к объему и производительности СХД. Построение соответствующих систем хранения и преодоление возникающих барьеров должно являться одной из ключевых задач современной суперкомпьютерной отрасли.

Существующие решения в области хранения данных

Стандартные

С точки зрения «обычных пользователей», бизнеса, наиболее удобными представляются "коробочные" версии высокопроизводительных систем хранения данных, представляющие собой настроенный и готовый к работе программно-аппаратный комплекс. Все ведущие поставщики ИТ-решений (IBM, HP, Oracle и другие) имеют в своих продуктовых линейках оригинальные либо заимствованные комплексы хранения данных. Существует и ряд специализированных компаний, которые успешно поставляют подобные хранилища «под ключ», такие как EMC. Приведем несколько лишь несколько примеров

К примеру, компания Dell Terascala [14] предоставляет высокопроизводительные решения в области хранения, стандартная конфигурация которых может предоставлять емкость 768TB для пользовательских данных, под управлением высокопроизводительной файловой системы Lustre и с поддержкой на 3 года. Другой лидирующей компанией на рынке США в области предоставления и поддержки высокопроизводительных систем хранения является компания Xyratex [15], которая также предоставляет высокопроизводительные хранилища с параллельным доступом. На российском рынке также существует достаточно широкий спектр предложений.

HPC

Для задач, находящихся ближе к переднему краю науки используются более сложные, единичные разработки, которые представляют собой сложную совокупность инженерных решений и программных продуктов.

Одним из способов оценить состояние суперкомпьютерного рынка является изучение статистики, предоставляемой рейтингом ТОП-500 [16]. ТОП-500 - это поддерживаемый в актуальном состоянии с 1993 года список самых мощных высокопроизводительных компьютеров в мире. Применительно к рассматриваемой теме, мы можем при помощи данного списка получить достаточно адекватный список СХД, обладающих достаточной надежностью, масштабируемостью и производительностью для использования на ведущих суперкомпьютерах мира. Изучение технических характеристик СХД, представленных на ведущих суперкомпьютерах мира, может позволить оценить тенденции развития систем хранения, отметить основные проблемы и намеченные пути их решения.

Согласно последнему рейтингу ТОП-500, который вышел в июне 2011 года, самой высокопроизводительной машиной мира является японский компьютер K computer. Несмотря на то, что этот компьютер уже лидирует в списке ТОП-500, согласно планам он будет окончательно сдан в эксплуатацию только в 2012 году.

Для данного суперкомпьютера разрабатывается система сверхвысокой масштабируемости FEFS [17]. Оперативная память одного узла K компьютера более 1PB, всего узлов планируется более 80,000. Предполагаемая файловая система должна обладать экстремально большой емкостью (от 100 PB до 1 EB), значительным количеством клиентом (100k~1M) и серверов (1k~10k)

Предполагаемые характеристики СХД:

- Пропускная способность одиночного потока (~GB/s) Параллельного ввода-вывода (~TB/s).
- Сокращенное время ожидания открытия файла (~10k ops).
- Всегда доступный файловый сервис, даже если какая-то часть системы сломана/недоступна.

Предполагаемая файловая система отражает планы и перспективы в направлении построения высокопроизводительных систем хранения для суперкомпьютеров в эру ОБОД приложений. Характеристики уже реализованных систем хранений на остальных установках первой десятки отражают реальное состояние СХД на сегодняшний день.

Рассмотрим таблицы, отражающие состояние СХД на ведущих суперкомпьютерах мира в 2011, 2006 и 2001 годах и проанализируем полученные сведения.

	Имя	Объем	Пропускная способность	Файловая система	Дополнительная информация
1	K computer	*(от 100 PB до 1 EB) ожидается*	*(~GB/s) Параллельного ввода-вывода (~TB/s). ожидается*	FEFS	Япония
2	Tianhe-1A [18]	1PB (2 PB по некоторым данным)		Lustre	Китай
3	Jaguar *[19]*	10 Петабайт	240 гигабайт/секунду	Spider (Lustre extension)	США
4	Nebulae	-	-	-	Китай
5	TSUBAME2.0 [20]	15 PB, иерархическое		7.13PB (Lustre + NFS Home)	Япония Дополнительно доступно 8 PB СХД на магнитных лентах
6	Cielo - Cray XE6 [21]	10 PB (в разработке)	160 GB/sec (в разработке)	PANASAS (в разработке)	США
7	Pleiades [22]	Всего доступно 6.9 PB total		7 файловых систем	США

				Lustre	
8	Hopper [23]	2 PB рабочей памяти	35 GB/sec	Lustre	США Дополнительно доступны все глобальные файловые системы NERSC, к примеру HPSS на 59 PB.
9	Tera-100 [24]	20PB	500GB/s	Lustre	Франция
10	Roadrunner [25]	2PB	~60GB/s	PANASAS	США
11	Kraken XT5 [26]	3.3 PB		NFS, HPSS	США
12	JUGENE [27]	5.3 PB	66 GB/s	GPFS	Германия
13	Lomonosov [28]	500 TB + 300 TB + 1 PB			Россия Трехуровневая СХД, включающая 500 TB T-Platforms ReadyStorage SAN, 300TB NAS storage и 1 PB на магнитных лентах
14	BlueGene/L [29]	1,89PB			США, содержит 1,024 Gb/s соединений с глобальной файловой системой
15	Intrepid [30]	~8PB	35 GB/s	GPFS	США

Июнь, 2011

	Имя	Объем	Пропускная способность	Файловая система	Дополнительная информация
1	BlueGene/L - eServer Blue Gene Solution				США
2	BGW - eServer Blue Gene Solution [31]	60 TB		GPFS	США Дополнительно используется 500 TB IBM 3494 на магнитных лентах
3	ASC Purple	1.6 PB (2 PB на	102 GB/s		США, эта система показывала

	[32]	2007 г.)			высокую пропускную способность, и позволила преодолеть так называемый "гигабайтовый барьер", выражающийся в неспособности интерконнекта большого суперкомпьютера "насытить" процессор данными
4	Columbia [33]	650 TB RAID storage			США; Дополнительно 10 PB на магнитных лентах
5	Tera-10 [34]	1 PB	100GB/s	Lustre	Франция
6	Thunderbird [35]	120 TB 50 TB	6.0 GB/s 4.0 GB/s	Lustre PANASAS	США, две файловые системы показаны в двух строках
7	TSUBAME Grid Cluster [36]	1PB (2007)	8 GB/sec	Lustre	Япония, первая промышленная система объединившая программный RAID Linux и Lustre.
8	JUBL	-	-	-	Германия
9	Red Storm [37]	340 TB, 1753 TB к 2008	Цель - 50.0 GB/s для каждого цвета	Lustre	США
10	Earth-Simulator	240 TB HDD RAID			Япония Иерархическое хранилище, 1.5 PB кассетных накопителей на магнитных лентах
11	MareNostrum [38]	280 TB			Испания, "самый красивый суперкомпьютер мира"
12	Stella				Нидерланды
13	Jaguar - Cray XT3 [39]	600 TB		Lustre	США
14	Thunder [40]	200TB	6.4 GB/s	Lustre	США

15	Blue Protein				Япония

Июнь 2006

	Name	Volume	Bandwidth	FS	Additional Info
1	ASCI White [41]	160 TB	-	GPFS	США
2	SP Power3	-	-	-	США
3	ASCI Red [42]	12.5TB RAID			США, Дополнительно было хранилище на магнитных лентах
4	ASCI Blue-Pacific SST [43]	62.5 TB - RAID5 0 глобальная файловая система; 17 TB – локальные диски	6.6 GB/s – глобальная; 11 GB/s – локальная файловая система.	GPFS	США, иерархическое хранилище, HDD на узлах.
5	SR8000/MPP				Япония, для вычислений с высокой точностью
6	ASCI Blue Mountain [44]	76 TB			США

Июнь 2001

Проанализируем полученные таблицы. Как известно, в базовые сведения, которые сообщаются в TOP-500 о каждом суперкомпьютере, информация о конфигурации системы хранения не входит, что согласуется с природой теста LINPACK, результаты которого не зависят от СХД. В связи с этим в заполнении таблиц есть пробелы, т.к. производители некоторых установок не публикуют данную информацию.

Тем не менее, в первую очередь следует отметить, что масштабы систем хранения претерпели не столь колоссальные изменения за 5 лет с 2006-2011. В 2006 году в первой десятке суперкомпьютеров лидирующим была СХД суперкомпьютера ASC Purple, которая обладала объемом в 1.6 PB и пропускной способностью в 102 GB/s (см. таблицу). Между тем, в 2011 году в первой десятке ведущих суперкомпьютеров мира уверенно держит место китайский суперкомпьютер Tianhe-1A с СХД достигающей, по различным данным, размера от 1 до 2 PB, т.е. сравнимую с системой хранения суперкомпьютера Purple. В первой десятке также можно

наблюдать суперкомпьютеры с пропускной способностью ввода/вывода не достигающей 100 GB/s (Hopper, Roadrunner -- из тех, про которые эти данные доступны), хотя этот барьер был также преодолен в 2006 году. Самыми лучшими характеристиками из первой десятки TOP-500 обладает система хранения суперкомпьютера Tera-100 (20 PB - объем, 500 GB/s - пропускная способность, т.е. в 12.5 раз больше объем, в 4.9 раз больше пропускная способность, чем у лучшего хранилища в 2001 году).

Для сравнения, теоретическая пиковая производительность с июня 2006 года изменилась с (18.20-280.60 TFLOPS) до (557.06 - 8773.63 TFLOPS). (первое значение - минимальная теоретическая пиковая производительность системы из 15, второе - максимальная, из ТОП-500), т.е. лучшая пиковая производительность увеличилась в 31 раз.

Кроме того, нельзя не отметить принципиальный разброс в объемах систем хранения за 2006 год (от 0,060 PB до 1.6 PB). В рейтинге за 2011 год разброс менее принципиален (от 1 PB до 20 PB), все системы (про которые доступна информация) вошедшие в первую десятку обладают системой хранения с объемом, превышающим 1 PB. Наметившаяся тенденция к выравниванию характеристик систем хранения показывает, что наличие адекватной системы хранения становится все более важным для современного суперкомпьютера. Следует также отметить, что суперкомпьютерные центры более развитых стран – США, стран Европы – обладают значительно превосходящими емкостями хранения по сравнению с машинами из Китая, хотя последние и могут занимать более высокое положение в рейтинге Linpack.

Тем не менее, заметный рост масштабов систем хранения суперЭВМ значительно ниже роста вычислительных мощностей. Сложно однозначно назвать причину, можно лишь сформулировать ряд предположений:

1) При построении рейтинга ТОП-500 не учитываются характеристики системы хранения. Тем не менее, именно рейтинг ТОП-500 имеет ключевое значение в мире высокопроизводительных вычислений, и представляет наибольший экономический и даже политический интерес для производителей, реальных пользователей.

2) Особенно актуальными задачи с интенсивной работой с данными стали именно сейчас. Это связано как с улучшением характеристик приборов наблюдения, получением данных с более высоким временным и пространственным разрешением, так и с накоплением архивов цифровых данных наблюдений, архивов данных от социальных сетей и экономических баз данных в беспрецедентных ранее масштабах.

3) Влияние на рост масштабов систем хранения могут оказывать технологические проблемы. Это и проблемы в области отказоустойчивости, надежности, обеспечении надлежащей пропускной способности, особенно в области системного ПО, включая файловые системы. В частности,

лидирующую позицию на сегодняшний день среди СХД для суперкомпьютеров занимает ФС Lustre. Между тем увеличение масштабируемости Lustre обходится в миллионы долларов и годы разработки.

Exascale

Также исследование возможностей увеличения масштабируемости, пропускной способности и надежности данных ведется в рамках инициатив по созданию вычислительного кластера экса-класса.

Roger Haskin из исследовательской группы IBM General Parallel File System предполагает, что увеличение масштабов суперкомпьютера до Exascale завершит извлечение файловой системы и хранилища из суперкомпьютера, т.е. система хранения будет существовать отдельно, аналогично файловому серверу, соединенная с вычислительным суперкомпьютером при помощи высокопроизводительных коммутационных решений. Он считает, что встроенные узлы ввода-вывода не предоставляют удовлетворительного объема памяти для оперирования данными и обладают рядом других недостатков.[45]

Известна концепция, что на высокомасштабируемых суперкомпьютерах будет применяться иерархическая система хранения. Одной из наиболее интересных разработок в этом направлении является файловая система Colibri, разрабатываемая под руководством Peter-a Braam-a в компании Xyratex ([46]). В системе предполагается промежуточный уровень - прокси, для быстрого сохранения большого объема данных, который будет предоставлять начальную пропускную способность и нижний уровень, для традиционного хранения, предоставляющий необходимый объем. Предполагается, что уровень прокси будет состоять из высокоростных твердотельных накопителей передового уровня технологии, а нижний уровень из более традиционных дисков. Идея буферизации данных (предоставления промежуточного аппаратного слоя между оперативной памятью и хранилищем) также разрабатывалась исследователями из национальной лаборатории США Аргон [47].

Другую позицию представляют исследователи из университета Токио, которые предполагают, что независимые системы ввода-вывода не демонстрируют надлежащую масштабируемость. Они предполагают, что использование развивающихся устройств хранения, таких как solid-state disks (SSDs) или Storage Class Memories (SCM) перспективны для организации ввода-вывода, увеличения производительности и оптимизации энергопотребления. Базируясь на данных технологических изменениях, японские ученые предлагают исследовать возможности активного хранения данных, уменьшения нагрузки на сохранение метаданных, анализ, организацию и перераспределение данных. [48]

Наибольшие технологические проблемы связаны не с разработкой аппаратной базы, а с изменением концепций программного обеспечения, для поддержки беспрецедентного уровеня масштабируемости. Вышеупомянутая файловая система Colibri обещает переопределить стандартные парадигмы хранения данных. В системе будет использоваться концептуально

отличная модель данных – Модель хранения объектов. Основную единицу будет представлять из себя объект-"контейнер", аналогичный Логическому Тому, но обладающий дополнительной операцией "вложение". Предполагается, что контейнеры можно будет вкладывать один в другой (вырожденный случай - перемещать) без разбора содержания. Такой подход позволяет значительно снизить накладные расходы на работу с метаданными и делает концепцию иерархического хранилища очень эффективной. База данных размещения контейнеров, предполагается более абстрактной, чем в текущих системах, базирующейся не на таблицах, а на формулах-зависимостях.

В системе предполагается вести журнал ошибок транзакций, который будет хранить информацию о всех совершенных действиях, для предоставления возможности отладки и мониторинга. Также исследователями предлагаются идеи об использовании опыта торрент-систем для повышения эффективности чтения и методы интеллектуального кэширования (использование опыта предыдущего запуска вычисляемой задачи и соответствующее перемещение данных, которые могут потребоваться в "быструю" память - твердотельный прокси слой)

Облачные вычисления

ОБОД приложения могут быть реализованы и в модели облачных вычислений. К традиционным преимуществам относятся перенос больших начальных затрат на покупку и поддержание дорогостоящего оборудования и организации центра данных, на «плоскую» систему оплаты услуг облачных инфраструктур. Когда речь идет о долгосрочном хранении данных и проведении различных исследований над сырыми данными в какой-то конкретной области науки, может идти речь о создании специализированного облака, к которому будут иметь доступ профильные специалисты.

Соответственно, с возросшими требованиями к разрешению, объемам и обработке данных, повышаются требования и к предоставляемой инфраструктуре системы хранения, предоставляемой облачной системой. В рамках этих требований модифицируются и создаются новые облачные платформы. Одним из примеров создающихся облачных платформ может являться платформа VISION Cloud [49], программа по созданию которой составлена с октября 2010 по сентябрь 2013. Цель данного проекта - создание мощной инфраструктуры для предоставления надежных и эффективных ОБОД сервисов хранения, упростить сближение информационных и коммуникационных технологий, СМИ и телекоммуникаций. В рамках данного проекта развивается более абстрагированная модель системы хранения, чем традиционные файловые системы, схожая с описанной моделью файловой системы Colibri.

Полный обзор облачных проектов выходит за рамки рассмотрения данной статьи, но это тоже развивающееся направление, включающее в себя объединение суперкомпьютеров, создание вычислительных сетей и масштабных центров данных.

Перспективы

Наиболее перспективным представляется развитие иерархических систем хранения данных с использованием твердотельных (SSD) дисков в качестве одного из уровней хранения. По ряду факторов, включая стоимость, производительность и надежность, SSD диски еще не могут полностью заменить HDD диски в высокопроизводительных СХД. Тем не менее, SSD диски могут и должны занять соответствующее место в иерархии хранения.

Эффективному использованию SSD дисков в высокопроизводительных СХД посвящен ряд проектов. К примеру, в статье [50] представлена комбинированная система хранения с SSD и HDD дисками, улучшенной производительностью, в проекте [51] показаны перспективы использования SSD дисков для хранения контрольных точек.

Colibri – уже названный выше проект, который подразумевает использование SSD в качестве промежуточного слоя в иерархии хранения. Как показывает исследование рейтинга ТОП-500, иерархические СХД, использующие магнитные носители в качестве одного из уровней хранения популярны и сейчас. Добавление нового уровня SSD может позволить повысить энергоэффективность и производительность СХД.

В ряде исследований можно заметить наметившиеся тенденции к приближению части вычислений к местам хранения данных за счет реализации технологий активного хранения. Такие предложения в рамках экса-исследований предлагают исследователи из университета Токио. Более детально эта мысль была рассмотрена в презентации [52], под названием "Киберкирпичи", по мотивам "активных дисков" Jim’a Gray. Для построения «кирпича» предлагается использовать материнскую плату Zotac Atom/ION, двухъядерный процессор Atom и 7.7 ТВ на SSD накопителях.

Достоинствами данной системы является низкое энергопотребление и способность выполнить часть операций по обработке данных непосредственно в пределах данного блока.

В ИПС РАН также проводились работы по исследованию возможностей активного хранения с использованием ФС Lustre, был получен прирост производительности [53].

Выводы

Сделанная выборка ОБОД приложений и определяемых ими требований к высокопроизводительным СХД позволяет предположить, что современных мощностей хранения недостаточно для удовлетворения запросов развивающейся науки. Количество данных возрастает быстрее, чем современные мощности хранения могут позволить сохранить и поддержать, в связи с чем приходится применять различные техники сжатия, фильтрации или просто удаления уже исследованных данных о проведенных экспериментах, что порождает риск потери ценной информации, которая могла бы пригодиться для дальнейших исследований.

Многие проекты, работающие со значительным объемом данных, называют цифры в сотни петабайт данных, в некоторых запросы доходят до эксабайта.

Увеличение объемов СХД приводит к проблемам надежности, производительности, изменению концепций работы с данными и метаданными.

Высокопроизводительные системы хранения данных, способные обеспечить пропускную способность выше 100 GB/s и объем более 1 PB, должны в ближайшем будущем войти в нашу жизнь как "стандартные" решения. Подобные мощности и объемы могут потребоваться для проведения маркетинговых исследований, создания фильмов, отслеживания социальных движений и т.п.

Для задач, находящихся у переднего края науки подобных мощностей уже недостаточно. Для К компьютера исследователи ставят себя цели в сотни PB, разработчики эксафлопсного проекта также называют цифры от 500-1000 PB с пропускной способностью 30-60 TB/сек.

Тем не менее, ведущие системы из актуального рейтинга TOP-500 сегодня обладают объемом систем хранения до 20 PB, при этом в первой десятке можно увидеть системы с хранилищем не превышающем 2 PB.

Основные проблемы остаются в области системного ПО, которое нуждается в увеличении пределов масштабирования, изменении ряда концепций, поддержки новых архитектурных решений. Ряд перспективных проектов в этом направлении позволяет предположить, что в ближайшем будущем большую популярность получат параллельные файловые системы с объектной моделью хранения данных. Многие исследователи видят перспективным развитие технологий активного хранения, выполнения хотя бы части операций по предобработке сырых данных непосредственно на узлах хранения, используя доступные вычислительные мощности.

Таким образом, можно выделить три наиболее перспективные тенденции в области высокопроизводительного хранения данных:

1. Изменение модели хранения данных (Объектная модель, приближение метаданных к данным, абстрагирование таблиц размещения)

2. Иерархическая система хранения данных (с уровнем SSD-накопителей)

3. Использование новых концепций (активное хранение, приближение части вычислений к местам хранения данных)

Вероятно также, что развитию высокопроизводительных СХД могло бы способствовать создание рейтинга, в чем-то аналогичного Graph 500, позволяющего сравнивать между собой высокопроизводительные системы хранения и анализа данных.

Литература

1. McKinsey & Company: *Big data: The next frontier for innovation, competition, and productivity*, URL: http://www.mckinsey.com/mgi/publications/big_data/pdfs/MGI_big_data_full_report.pdf
2. *IBM's Top Storage Predictions for 2011*, Январь 2011, StorageNewsletter.com, URL: http://www.storagenewsletter.com/news/miscellaneous/ibm-top-storage-predictions-for-2011
3. *Avatar takes 1 petabyte of storage space*, Январь, 2010, http://www.devilsduke.com/avatar-takes-1-petabyte-of-storage-space/608/
4. *Facebook has the world's largest Hadoop cluster!*, май 2010, URL: http://hadoopblog.blogspot.com/2010/05/facebook-has-worlds-largest-hadoop.html
5. *The Fourth Paradigm: Data-Intensive Scientific Discovery,* 2009, URL: http://research.microsoft.com/en-us/collaboration/fourthparadigm
6. Goble, C. and De Roure, D.: *The impact of workflow tools on data-centric research*. In: Data Intensive Computing: The Fourth Paradigm of Scientific Discovery, 2009.
7. *Data Intensive Computing*, http://dicomputing.pnnl.gov/
8. *The Graph 500 list*, URL: http://www.graph500.org/index.html
9. Loek Essers: *CERN pushes storage limits as it probes secrets of universe*, URL: http://news.idg.no/cw/art.cfm?id=FF726AD5-1A64-6A71-CE987454D9028BDF
10. University of Hawaii: *World's Largest Digital Camera Installed on Maui Telescope*, август, 2007, URL: http://www.ifa.hawaii.edu/info/press-releases/GPC/gigapixel_camera-8-07.html
11. Duncan Brown, Syracuse University: *Computational Challenges in Gravitational Wave Astronomy*, URL: http://www.psc.edu/data-analytics/proceedings/BrownSlides.pdf
12. Hacker T. J., Eigenmann, R., Irfanoglu, A., Pujol, S., Rathje, E., Catlin, A., Bahchi, S.: *Developing an Effective Cyberinfrastructure for Earthquake Engineering: The NEEShub*, In IEEE Computing in Science & Engineering, 2011 (Invited Paper.)
13. Arthur W. Wetzel, Greg Hood: *Connectomics: Challenges in Reconstructing Neural Circuitry from Massive Serial Section Electron Microscopy Datasets;*Data-Intensive Analysis, Analytics and Informatics TeraGrid/Blue Waters Symposium, Апрель 2011, URL: http://www.psc.edu/data-analytics/proceedings/WetzelSlides.pdf
14. *Dell | Terascala HPC Storage Solution*, URL:http://www.terascala.com/dell-terascala-hss.html
15. Xyratex - *Advancing Digital* Storage *Innovation*, URL: http://www.xyratex.com/
16. *TOP-500 supercomputer sites*, URL: http://top500.org/
17. Shinji Sumimoto: *An Overview of Fujitsu,,s Lustre Based File System*, Apr.12 2011, URL: http://www.olcf.ornl.gov/wp-content/events/lug2011/4-12-2011/230-300_Shinji_Sumimoto_LUG2011-FJ-20110407-pub.pdf
18. *Tianhe-1 Pflop Supercomputer*, URL: http://nscc-tj.gov.cn/en/show.asp?id=191
19. Arthur S. Bland, Ricky A. Kendall, Douglas B. Kothe, James H. Rogers, Galen M. Shipman, Oak Ridge National Laboratory: *Jaguar: The World's Most Powerful Computer*, CUG 2009 Proceedings, URL: http://www.nccs.gov/wp-content/uploads/2010/01/Bland-Jaguar-Paper.pdf

20. Satoshi Matsuoka: *TSUBAME2.0: A Tiny and Greenest PetaflopsSupercomputer*, Nov 2010, URL: http://www.nvidia.com/content/PDF/sc_2010/theater/Matsuoka_SC10.pdf
21. Garth Gibson: *Data Systems @ Scale*, Carnegie Mellon University, 9 февраля 2011, URL: http://www.cs.cmu.edu/~pll/CNOSSG/Gibson-CNOssg-Feb9.pdf
22. *Pleiades Supercomputer*, NAS Division Website, URL: http://www.nas.nasa.gov/hecc/resources/pleiades.html
23. *Hopper*, National Energy Research Scientific Computing Center (NERSC), URL: http://www.nersc.gov/users/computational-systems/hopper/
24. Peter Sayer: *Bull Bills Tera 100 as Europe's First Petaflop Computer*, IDG News, Май 2010, URL: http://www.pcworld.com/businesscenter/article/197454/bull_bills_tera_100_as_europes_first_petaflop_computer.html
25. Brent Welch*: Exascale Distributed File Systems*, MSST, Май, 2010, URL: http://storageconference.org/2010/Presentations/MSST/8.Welch.pdf
26. *Kraken*, National Institute for Computational Sciences (NICS), URL: http://www.nics.tennessee.edu/computing-resources/kraken
27. N. Attig, F. Berberich, U. Detert, N. Eicker, T. Eickermann, P. Gibbon, W. Gurich, W. Homberg, A. Illich, S. Rinke, M. Stephan, K. Wolkersdorfer, and T. Lippert: *Entering the petaflop-era - new developments in supercomputing*. In G. Munster, D. Wolf, and M. Kremer, editors, NIC Symposium 2010, volume 3, pages 1-12. IAS Series, 2010
28. *MSU SUPERCOMPUTERS: "LOMONOSOV*", URL: http://hpc.msu.ru/?q=node/59
29. *BlueGene/L Configuration*, Lawrence Livermore National Laboratory, URL: https://asc.llnl.gov/computing_resources/bluegenel/configuration.html
30. Jing Fu, Ning Liu: *Scalable Parallel I/O Alternatives for Massively Parallel Partitioned Solver Systems*, URL: http:// cmes.colorado.edu/courses/hpc/ipdps-lspp-parallel-io-04-23-2010-1.ppt
31. *BGW*, TOP-500 supercomputer sites, URL: http://top500.org/system/7466
32. Clint Boulton: *IBM: The Power of Purple*, Март, 2006, URL: http://www.internetnews.com/ent-news/article.php/3590236/IBM-The-Power-of-Purple.htm
33. *Columbia*, TOP-500 supercomputer sites, URL: http://top500.org/system/7288
34. Peter Bojanic: *LUSTRE ROADMAP and FUTURE PLANS*, Sun HPC Consortium, Июнь, 2008, URL: http://www.hpcuserforum.com/presentations/Tucson/SUN%20%20Lustre_Update-080615.pdf
35. Jerry D. Smith II*: Thunderbird Capacity Computing System*, Sandia National Laboratories, May 3, 2006, URL: http://www.linuxclustersinstitute.org/conferences/archive/2006/PDF/ThunderbirdUpdate.pdf
36. Syuuichi Ihara: *TOKYO TECH TSUBAME GRID STORAGE IMPLEMENTATION*, Sun BluePrints™ On-Line, May 2007, Part No 820-2187-10, Revision 1.0, 5/22/07, URL: http://www.filibeto.org/sun/lib/blueprints/820-2187.pdf

37. *Red Storm upgrade lifts Sandia supercomputer to 2nd in world, but 1st in scalability, say researchers*, ноябрь, 2006, URL: https://share.sandia.gov/news/resources/releases/2006/red-storm.html
38. *MareNostrum*, TOP-500 supercomputer sites, URL: http://top500.org/system/8242
39. *Jaguar*, TOP-500 supercomputer sites, URL: http://top500.org/system/7938
40. Robin Goldstone: *The Roar of Thunder: LLNL Goes Itanium in a Big Way*, Lawrence Livermore National Laboratory, Presented to Gelato.org, Май, 2004, UCRL-PRES-204277, URL: http://www.gelato.org/pdf/Illinois/gelato_IL2004_goldstone_llnl.pdf
41. *ASCI White*, URL: https://computation.llnl.gov/casc/sc2001_fliers/ASCI_White/ASCI_White01.html
42. *ASCI Red*, TOP-500 supercomputer sites, URL: http://www.top500.org/system/4428
43. Mark Seager: *An ASCI Terascale Simulation Environment Implementation*, UCRL-JC-134806 PREPRINT, Mannheim Supercomputer '99 Conference, June 11, 1999, URL: https://e-reports-ext.llnl.gov/pdf/235862.pdf
44. *Overview of the Advanced Simulation and Computing Program (ASCI)*, UKHEC, URL: http://www.ukhec.ac.uk/publications/reports/asci.pdf
45. Roger Haskin: *Exascale Storage Challenges*, 2010, IBM Corp, URL: http://institute.lanl.gov/hec-fsio/conferences/2010/presentations/day3/Haskin-HECFSIO-2010-ExascaleChallenges.pdf
46. Peter Braam: *Exascale File Systems, Scalability in ClusterStor's Colibri System*, 2010, URL: http://www.teratec.eu/forum_2010/Presentations/A5_Braam_ClusterStor_Forum_Teratec_2010.pdf
47. Rob Ross: *Storage in an Exascale World*, Argonne National Laboratory, URL: http://storageconference.org/2010/Presentations/SNAPI/1.Ross.pdf
48. Yutaka Ishikawa: *Towards Exascale File I/O*, University of Tokyo, Japan, 2009/05/21, http://www.exascale.org/mediawiki/images/6/65/ExascaleFile-io-ishikawa071309.pdf
49. Mirko Lorenz: *Vision Cloud: The Fact Sheet*, 20.12.2010, URL: http://www.visioncloud.eu/content.php?s=30,47
50. Youngjae Kim, Aayush Gupta, Bhuvan Urgaonkar, Piotr Berman, and Anand Sivasubramaniam: HybridStore*: A Cost-Efficient, High-Performance Storage System Combining SSDs and HDDs*, Proceedings of the the IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS), Singapore, July 2011.
51. N. Kämmer, S. Gerhold, A. Weggerle, C. Himpel, P. Schulthess: *Pageserver: High-Performance SSD-based Checkpointing of Transactional Distributed Memory*, Proceedings of the 2nd International Conference on Computing Engineering and Applications (ICCEA 2010), Bali, Indonesia, 2010.
52. Alex Szalay: *Extreme Data-Intensive Computing*, The Johns Hopkins University, 19 May 2011, URL: http://salsahpc.indiana.edu/tutorial/slides/0726/szalay-bigdata-2010.pdf

53. Шевчук Е. В., Тютляева Е. О., Московский А. А. 2009. *Система активного хранения данных на базе библиотеки динамического распараллеливания TSim*. // Научный сервис в сети Интернет: масштабируемость, параллельность, эффективность. Труды Всероссийской научной конференции, 21-26 сентября 2009 г. Новороссийск, — М.: Изд-во МГУ им. М.В. Ломоносова, 2009 с. 226—230 (CD) ISBN 978-5-211-05697-8
54. DKRZ brochure (2009) "The power to understand: Supercomputing for Climate System Science"
55. http://www.graph500.org/june2011.html, позиция 7.

Сведения об авторах

Рус:

1. Тютляева Екатерина Олеговна
2. Учреждение Российской академии наук Институт программных систем им. А.К. Айламазяна РАН, инженер-программист
3. 2009, НОУ Высшего профессионального образования ИНСТИТУТ ПРОГРАММНЫХ СИСТЕМ – «УНИВЕРСИТЕТ ГОРОДА ПЕРЕСЛАВЛЯ» имени А.К. Айламазяна
4. Нет
5. 12
6. Системы хранения данных, высокопроизводительный ввод-вывод, отказоустойчивость
7. ordi@xgl.pereslavl.ru, +7(960)5399351

1. Московский Александр Александрович
2. Директор по науке, ЗАО "РСК СКИФ"
3. 1997, МГУ им. М.В. Ломоносова, Химический факультет
4. К.х.н.
5. 20
6. молекулярное моделирование, высокопроизводительные вычисления
7. moskov@rsc-skif.ru, +7(916)5578382

Анг:

1. Tyutlyaeva Ekaterina
2. Program System Institute of RAS, engineer-programmer
3. 2009, Aylamazyan University of Pereslavl
4. -
5. 12
6. Storage systems, high-performance I/O, fault-tolerance

7. ordi@xgl.pereslavl.ru, +7(960)5399351

1. Moskovsky Alexander
2. Science director, ZAO “RSK SKIF”
3. 1997, MSU
4. phD
5. 20
6. Molecular modelling, HPC
7. moskov@rsc-skif.ru, +7(916)5578382