ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ № ?, 2014, с. ??–??

УДК 004.62

Е. О. Тютляева, М. М. Тютляев

Системы хранения данных лидирующих суперкомпьютеров

Аннотация. В статье рассматриваются основные тенденции в области организации высокопроизводительных систем хранения данных для ведущих суперкомпьютеров мира, проводится сравнительный анализ полученных характеристик. Исследуются ключевые моменты как в области программного, так и аппаратного обеспечения соответствующих систем и формулируются основные проблемы, лежащие перед специалистами в области хранения и обработки данных. На основании проанализированных фактов и изученных документов составлен краткий прогноз основных характеристик систем хранения, которые будут занимать лидирующие позиции в ближайшем будущем.

Ключевые слова и фразы: Системы хранения данных, Большие данные, ТОП-500, Распределенное хранение данных, многоуровневое хранилище.

Введение

Большие данные в настоящее время используются во всех областях человеческой деятельности, от фундаментальных наук до организации эффективных рекламных кампаний. Стоит привести ряд фактов, чтобы представить масштаб проблемы, которую нужно решать специалистам в области хранения данных прямо сейчас. Всего несколько лет назад количество электронных устройств, имеющих выход в сеть, сравнялось с численностью населения Земли [1]. Но уже к 2015 или 2016 году количество устройств будет превышать численность населения Земли в два раза. В тех же источниках утверждается, что к 2015 году потребуется пять лет, чтоб посмотреть все видео, которое проходит через сети IP за одну секунду. Известная популярная сеть Фейсбук собирает 500 ТБ данных каждый день. За последние

два года человечество собрало больше данных, чем за всю предыдущую историю [2]. Исследования утверждают, что к 2015 году количество вакантных мест для специалистов по работе с данными и аналитиков достигнет 4.4 миллионов, и только 1/3 этих мест будет занята [3].

Подобные факты доказывают, что скачок в области работы с данными не просто количественный, но уже и качественный. Современные исследования в области философии науки утверждают, что наступила новая эпоха в области получения научного знания, так называемая четвертая парадигма. Под этим термином подразумевается новый метод получения научного знания, основанный на технологиях сбора, анализа, визуализации и поиска закономерностей в больших массивах данных [4].

Следующие примеры [2] демонстрируют, насколько широко проводятся исследования, связанные с обработкой больших объемов данных:

- Помощник профессора управления Гарварда Кувин Куинн провел следующий эксперимент в области юриспруденции: 87 профессорам в области юриспруденции предложили предсказать решение Верховного Суда по рассмотренным делам за последний год. Все профессора отлично разбирались в юриспруденции и знали решение Верховного Суда в предшествующих аналогичных случаях. Их решение сравнивалось с решением, которое рассчитывала статистическая модель на основании доступных данных. Эксперимент продемонстрировал, что статистическая модель однозначно побеждает в производительности и точности решения как одного профессора, так и небольшие группы.
- В маркетинге анализ больших объемов данных позволяет разработать чрезвычайно эффективные рекомендательные сервисы и оценить клиентов. Известно, что большие корпорации, например Netflix и Amazon предлагают покупателям товары, базируясь на предпочтениях других покупателей. Более интересными примерами могут служить исследования, которые определяют, что женщина-клиент беременна, если она заинтересована в лосьонах для тела без отдушек, или исследование компаний предоставляющих кредитные карты, что показателем благонадежности клиента является приобретение им противоскользящих ковриков под мебель.
- В общественной сфере проводится огромное количество иссле-

дований, выявляющих взаимосвязь между условиями окружающей среды и здоровьем населения, предсказывающих увеличение преступности в определенном районе и т.п.

Очевидно, что в областях науки, которые мы традиционно привыкли связывать с обработкой данных, таких, к примеру, как астрономия и сейсмология, получение знания из больших объемов данных также позволяет совершить качественный скачок, за счет беспрецедентной точности наблюдения и увеличения вычислительных возможностей по обработке накопленных данных.

Рассмотренные примеры доказывают, что организация качественного процесса работы с данным становится первоочередным вопросом в области высокопроизводительных вычислений. «Сейчас существует исключительно мало практически значимых высокомасштабируемых приложений, которые НЕ работают интенсивно с данными»[1] . Указанная тенденция предъявляет особые требования к системам хранения данных, которые изменяют профиль области.

В настоящей статье будут рассмотрены основные тенденции в области организации высокопроизводительных систем хранения данных для ведущих суперкомпьютеров мира, будут выделены ключевые моменты как в области программного, так и аппаратного обеспечения соответствующих систем и сформулированы основные проблемы, лежащие перед специалистами в области хранения и обработки данных.

1. Анализ рейтинга ТОП-500

Одним из признанных средств для анализа профиля области является рейтинг ТОП-500, который позволяет оценить те системы хранения, которые используются лидерами в области высокопроизводительных вычислений. В 2011 году я представляла на Московском Суперкомпьютерном Форуме доклад, посвященный анализу основных тенденций хранения данных [5]. В июне 2011 года системы хранения, относящиеся к десяти ведущим установкам мира, имели характеристики, приведенные в таблице 1

[1] Alok Choudhary, IESP, Kobe, Japan, Apr

Таблица 1: Системы хранения на ведущих суперкомпьютерах за июнь, 2011

№	Имя	Компьютер	Объем	Проп. способность	Файловая система
1	K computer	SPARC64 VIIIfx 2.0GHz, Tofu interconnect	(100 PB – 1 EB expected)	GB/s (TB/s) expected	FEFS (Lustre)
2	Tianhe-1A	NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C	1-2 PB (depends on source)		Lustre
3	Jaguar	NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C	10 PB	240 GB/s	Spider (Lustre extension)
4	Nebulae	Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU	?	-	-
5	TSUBAME2.0	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows	15 PB, tiered		7.13PB (Lustre + NFS Home)
6	Cielo - Cray XE6	Cray XE6 8-core 2.4 GHz	10 PB (в разработке)	160 GB/sec (в разработке)	PANASAS (в разработке)

Продолжение таблицы 1

7	Pleiades	SGI Altix ICE 8200EX/8400EX, Xeon HT QC 3.0/Xeon 5570/5670 2.93 Ghz, Infiniband	6.9 PB	-	7 Lustre
8	Hopper	Cray XE6 12-core 2.1 GHz	2 PB	35 GB/s	Lustre
9	Tera-100	Bull bullx super-node S6010/S6030	20PB	500GB/s	Lustre
10	Roadrunner	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband	2PB	60GB/s	PANASAS

Актуальный рейтинг за июнь 2014 года демонстрирует следующие системы хранения (см. таблицу 2).

Таблица 2: Системы хранения на ведущих суперкомпьютерах за июнь, 2014

№	Имя	Компьютер	Объем	Проп. способность	Файловая система

Продолжение таблицы 2

1	Tianhe-2	TH-IVB-FEP Cluster, Xeon E5-2692 12C 2.2GHz, TH Express-2, Intel Xeon Phi	12.4 PB	750 GB/s	Lustre/ H2FS
2	Titan	Cray XK7 , Opteron 6274 16C 2.2GHz, Cray Gemini interconnect, NVIDIA K20x	10.5 PB	240 GB/s	Lustre
3	Sequoia	BlueGene/Q, United 55 PB Power BQC 16C 1.60 GHz, Custom Interconnect	55 PB	850 GB/s	Lustre
4	K computer	Fujitsu, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	40 PB	965 GB/s	Lustre
5	Mira	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	7.6 PB	88 GB/s	GPFS
6	Piz Daint	Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x	2.5 PB	138 GB/s	Lustre

Продолжение таблицы 2

7	Stampede	PowerEdge C8220, Xeon E5-2680 8C 2.7GHz, IB FDR, Intel Xeon Phi	14 PB	150 GB/s	Lustre/ H2FS
8	JUQUEEN	BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect	5.6 PB	33 GB/s	GPFS
9	Vulcan	BlueGene/Q, Power BQC 16C 1.600GHz, United States Custom Interconnect	55 PB	850 GB/s	Lustre
10	Cray XC30	Intel Xeon E5-2697v2 12C 2.7GHz, Aries interconnect	?	?	Lustre

Анализ изменений, которые произошли с системами хранения первой десятки самых высокопроизводительных установок мира, позволяет оценить наиболее популярные тенденции.

Можно отметить увеличение объема систем хранения на ведущих установках. В 2011 году лидирующая тройка по объему составляла 20 PB (Tera-100), 15 PB (Tsubame2.0) и 10 PB (Cielo, Jaguar). В июне 2014 года лидирующая тройка составляет 55 PB (Sequoia), 55 PB (Vulcan) и 40PB (K computer). 10 PB в 2011 году является третьим по значимости показателем в 2011, в то время как в 2014 году как минимум 6 систем из 10 уверенно превышают этот показатель. Максимальный показатель пропускной способности (965 GB/s,

K computer) увеличился почти в два раза (по сравнению с максимальным за 2011, 500 GB/s, Tera-100) и уверенно приближается к показателю в 1 TB/s. Интересно также проследить эволюцию системы хранения на K computer-е. В 2011 году было заявлено, что ожидается расширение системы хранения до объема в экзабайт данных и пропускной способности в 1 TB/s. В настоящее время пропускная способность (965 GB/s) действительно соответствует заявленной, в то время как доступный объем (40 PB) однозначно ниже ожидаемого на два порядка.

Впрочем, следует отметить, что с объемом систем хранения все не так однозначно. Во-первых, большинство установок имеет доступ к системам хранения данных на магнитных лентах. К примеру, с суперкомпьютера Stampede доступна 60PB архивная система TACC Ranch tape archival system [9]. Titan имеет доступ к высокопроизводительной системе хранения HPSS, которая предоставляет возможность долговременного хранения данных и содержит три магнитные библиотеки SL8500, каждая из которых состоит из 10000 картриджей [14]. Аналогично, доступ к архивной системе есть с установок sequoia и Vulcan [15].

Во вторых, ряд суперкомпьютеров из первой десятки разделяют систему хранения как инфраструктуру на различных уровнях, от организационного, до государственного. Так, K computer играет ключевую роль в государственной японской инфраструктуре HPCI, в которую также входит суперкомпьютер TSUBAME 2.5, занимающий тринадцатую строку рейтинга.

В области программного обеспечения лидирующую роль сохраняет файловая система Lustre и ее различные модификации. Также в первую десятку рейтинга вошли системы, использующие GPFS, и хотя по средним показателям они проигрывают системам, базирующимся на Lustre, они демонстрируют перспективные подходы к организации системы хранения, что может повысить их конкурентоспособность. Например, установка Mira использует оригинальную версию GPFS системы с решением, отсылающим нас к концепции “burst buffer”, которое будет более подробно рассмотрено ниже [16] Кластерная файловая система PANASAS, в то же время, окончательно вытеснена из первой десятки. Как отмечает в своем докладе Торбен Клинг Петерсен [17], PANASAS исчерпала свои возможности, к тому же она до сих пор базируется на RAID5, который уже признан неспособным поддержать современные запросы в области

хранения и защиты данных. Впрочем, в июне 2014 года компания panasas представила новую версию системы PanFS6.0 [28], базирующуюся на инновационных принципах, которые могут повысить ее конкурентоспособность в ближайшем будущем.

Рассмотрим подробнее наиболее интересные технологии, которые позволил выявить анализ рейтинга TOP-500.

2. Перспективные технологии

2.1. Системы управления данными

Как демонстрирует рейтинг ТОП-500, лидирующие позиции в области HPC файловых систем по прежнему занимает распределенная файловая система Lustre. Стоит отметить, что кроме масштабных систем хранения, перечисленных в рейтинге, Lustre также установлена на суперкомпьютере NCSA Bluewaters, который в рейтинг не входит. Система хранения на этом суперкомпьютере обладает замечательными характеристиками с емкостью в 24 PB и рекордной пропускной способностью в 1100 GB/S. Указанные характеристики однозначно позволяют назвать Lustre лидирующей параллельной файловой системой.

Несмотря на это, на уровне широкого круга пользователей высокопроизводительных кластеров Lustre и GPFS обладают сравнимой популярностью. Согласно результатам переписи Intersect360 за 2014 год [6], прогресс в области адаптации параллельных файловых систем проходит чрезвычайно медленно, наиболее решительные шаги наблюдаются в области академических и правительственных исследований. Большинство ПО, обеспечивающего управление системой хранения, (36%) предоставляется производителем систем хранения. Lustre и GPFS являются наиболее часто упоминаемыми системами с показателями в 19% и 17%. Перспективным направлением является создание киберинфраструктуры, объединяющей несколько ресурсов. Рассмотрим наиболее интересные системы, которые используются для организации таких киберинфраструктур, в частности уровня хранения, для описания которого иногда применяется термин DataGrid.

Японская киберинфраструктура HPCI базируется на распределенной файловой системе Gfarm (название произведено от Grid Data

Farm). Другим перспективным ПО является система управления данными на основе правил iRODS, которая используется в киберинфраструктуре iPlant, которая используется более чем десятью тысячами пользователей. Ключевой особенностью iRODS является возможность управления различными цифровыми объектами, сохраненными на множестве систем (NFS, HDFS, HPSS и т.д.). При этом сохраняется единое и целостное пространство имен, централизованная система метаданных, множество клиентов и API и возможность установки гибких настроек безопасности и управления данными (запросы через web, SQL, Hadoop, использование специализированных рабочих потоков и правил для обработки данных...). Кроме того, iRODS предоставляет удобные инструменты для создания резервных копий данных, что является критически важным параметром при современных масштабах систем хранения. Перечисленные возможности нацелены на предоставление максимального спектра инструментов при абстрагировании от деталей реализации, которые должны позволить специалистам максимально сосредоточиться на решении исследовательской задачи, а не на технических деталях работы с большими данными. Кроме инициативы iPlant IRODS используется в ряде академических организаций и научных репозиториев по всему миру. Статистика использования iRODS [24] позволяет оценить данную систему как одну из наиболее популярных на сегодняшний день систем для организации академических Data-грид киберинфраструктур.

2.2. Burst Buffer

Несмотря на очевидный достигнутый прогресс, перед создателями системы хранения по прежнему стоит ряд технологических проблем, которые нужно решить для создания соответствующей системы хранения для суперкомпьютера экса-класса. Ключевой из этих проблем является обеспечение такой производительности ввода-вывода, которая бы позволила сохранять промежуточные и окончательные результаты вычисления в адекватный период времени. Питер Браам при проектировании системы Colibri для хранилища экса-класса называл следующие ориентировочные показатели: желательно, чтобы ввод-вывод данных, полученных за час вычисления, занимал не более пяти минут. Минимизация времени ввода-вывода также необходима для уменьшения вероятности возникновения ошибки во время этой критической транзакции и для минимизации времени на сохранение контрольных точек. Согласно ведущим исследователям, для того,

чтобы обеспечить адекватную скорость сохранения контрольных точек, в ближайшем будущем системы должны будут предоставлять пропускную способность приблизительно равную 60 TiB/s [11, 12].

Одним из решений, которое позволит минимизировать время ввода/вывода для приложений и обеспечить надлежащую пропускную способность является так называемый "Burst Buffer". Этот термин был введен в использование Los Alamos National Labs' HPC division leader, Gary Grider, который рассматривает указанные буферы как значительный потенциал для суперкомпьютеров экса-класса [19]. Под этим термином понимается включение промежуточного уровня хранения , состоящего из устройств хранения, обеспечивающих высокую пропускную способность и сравнительно небольшой объем хранения, которые действуют как участок подготовки необходимых данных или как кэш с обратной записью для высокопроизводительных систем хранения [10]. Интересным подходам является интеграция таких буферов к узлам ввода-вывода, как часть I/O forwarding services. Исследование, проведенное с использованием суперкомпьютера Intrepid [18] однозначно демонстрирует, что использование burst buffer'а является эффективным решением как для повышения производительности стадии ввода-вывода, так и для сокращения нагрузки на внешнюю систему хранения. Таким образом, дополнительным преимуществом решения burst buffer является значительное сокращение расходов на организацию внешнего хранилища в связи со снижением требований к производительности ввода-вывода. Указанный подход однозначно выигрывает у традиционного подхода, заключающегося в обеспечении высокой пропускной способности для внешней системы хранения, так как полученная система будет использоваться на полную мощность лишь незначительную часть времени. (В работе [13], к примеру, рассмотрен низкий уровень загрузки суперкомпьютера IBM Blue Gene/P "Intrepid")

Авторы [10] утверждают, что вероятнее всего к 2020-му году burst buffer станут обязательным компонентом высокопроизводительных установок.

Уровень "Burst Buffer" планируется включить в такие суперкомпьютеры, как Trinity и NERSC-8: Cori, которые в настоящее время разрабатываются под началом компании Cray и являются одними из самых перспективных и ожидаемых на настоящий момент суперкомпьютеров. Ожидаемая пропускная способность burst buffer для суперкомпьютера Тринити находится в пределах 4.4-17.8 TB/s. Дан-

ные показатели связаны с ожидаемым средним временем до отказа задачи (jMTTI) в пределах 10-20 часов и общим размером оперативной памяти в 2-4 PB. Для суперкомпьютера Кори при объеме оперативной памяти в 1-2 PB соответствующая минимальная пропускная способность составляет от 2.2 до 8.9 TB/s. В настоящее время доступна спецификация сценариев работы Burst Buffer для данных суперкомпьютеров Trinity и Cori [20]. Она включает в себя описание первичного сценария для работы с контрольными точками и вспомогательных сценариев для кэширования данных, сохранения временных данных работающей задачи и работы в режиме анализа и визуализации данных. Согласно предложенным сценариям, контрольные точки записываются в burst Buffer с максимальной производительностью, после чего в фоновом режиме для каждой n-й контрольной точки (или через определенный промежуток времени) выполняется копирование данных контрольных точек в основную файловую систему с целью обеспечения достаточного объема доступного свободного места и обеспечения отказоустойчивости на случай возникновения ошибок на уровне Burst Buffer-а. Остальные сценарии описывают ситуации загрузки в burst Buffer конфигурационных файлов и объектных библиотек, которые используются параллельно несколькими узлами во время или до запуска приложения; использования буфера для обработки больших объемов данных, которые не помещаются в оперативную память; анализ и сравнение загружаемых данных с данными, полученными на предыдущем шаге. В документе также указаны рекомендации к реализации постраничных стратегий и потоковых алгоритмов, возможно, с использованием дополнительных вычислительных ресурсов присоединенных к burst Buffer-у для оптимизации работы с объемами данных, которые не помещаются в оперативную память. Согласно документу, создание уровня Burst Buffer должно быть экономически оправдано за счет увеличения эффективности обработки рабочих потоков, необходимое, чтобы обеспечивать надлежащую производительность при ожидаемом увеличенном количестве ошибок (включая ошибки на уровне Burst Buffer-а).

В настоящее время, как уже было отмечено, подобное решение реализовано на суперкомпьютере Mira из первой десятки TOP-500, в системе хранения которой используется дополнительный уровень хранения, обладающий следующими характеристиками:

- Обладает высокой пропускной способностью, но ограниченной емкостью, которая не позволяет хранить все данные постоянно

(аналогично кэш-памяти)
- Улучшает производительность для большинства типичных сценариев ввода-вывода
- Позволяет перемещать данные без прямого вмешательства пользователя.

2.3. Многоуровневое хранилище

С приходом концепции burst buffer и уровня SSD накопителей, как такового, стала перспективной архитектура многоуровневого хранилища, которая позволяет интегрировать в единую систему уровни SSD, различные традиционные диски и зачастую хранилище на магнитных лентах. Основная проблема, разумеется, лежит не в аппаратном соединении всех этих уровней в единую систему хранения, а в разработке программного обеспечения, которое позволит предоставить единообразный доступ ко всем уровням полученной системы и обеспечить автоматическое перемещение между уровнями согласно правилам, которые пользователь сможет настраивать согласно своим нуждам. Таким образом, необходим гибкий и функциональный уровень абстракции над слоями хранения, представленными различными производителями, различным аппаратными решениями и различными характеристиками.

Ряд ведущих коммерческих компаний уже представляет соответствующие разработки.

Например, компания Cray предоставляет конечное решение TAS (Tiered Adaptive Storage) [25], которое состоит из четырех базовых уровней. Нулевой уровень оптимизирован для организации ввода-вывода с высокой пропускной способностью и обычно представлен SSD накопителями, реже жесткими дисками. Первый уровень предназначен для хранения данных большую часть времени и состоит из эффективных жестких дисков или SSD. Наконец, второй и третий уровни предоставляют максимальную емкость, третий уровень максимально оптимизирован по стоимости и нацелен на долговременное хранение редео (никогда) не использующихся данных и чаще всего представлен магнитными накопителями. Для управления данными используется ряд инструментов, объединенных системой упрправления Versity, которые позволяют определять и настраивать различные стратегии хранения, которые обеспечивают миграцию данных между уровнями и создание резервных копий при достижении специфических указанных условий.

В свою очередь IBM предлагает решение Easy Tier®, которое автоматически перемещает часто используемые данные с традиционных дисков на SSD диски согласно алгоритмам, разработанным в IBM Research, которые вычисляют частоту доступа к данным и перемещают более «горячие» данные (к которым наиболее частый доступ) на SSD диски, а «остывшие» данные (частота доступа к которым снизилась) на традиционный уровень хранения.

2.4. SSD

Использование SSD накопителей в высокопроизводительных установках все еще находится на ранней стадии развития, согласно исследованию InterSect360. Только 14% систем, модифицированных после 2012 года используют SSD как минимум на нескольких узлах. При этом очень мало систем используют только SSD накопители, чаще всего SSD применяется для организации дополнительного слоя между памятью и традиционными жесткими дисками.

2.5. Хранилища на магнитных лентах

Стоит отметить, что несмотря на развитие технологий хранения, магнитные носители по прежнему предлагают лучшее соотношение цена/объем, кроме того, они обладают высокой надежностью и выгодны в обслуживании. По указанным причинам, они все еще популярны, особенно для хранения редко (никогда не) используемых данных. Согласно исследованию InterSect360 за 2013 год [7], 30% пользователей крупнейших высокопроизводительных установок используют магнитные хранилища для хранения архивных данных.

2.6. Объектное хранилище

По прежнему лидирующую позицию сохраняет концепция объектного хранения цифровых объектов, которая должна в ближайшем времени вытеснить файловые системы. Lustre является наиболее ярким примером файловой системы, которая базируется на объектах, а не на файловой иерархии. Каждый файл и директория в Lustre рассматривается как отдельный объект с определенными атрибутами, метаданные, включающие информацию о распределении файла в системе хранения, содержатся на сервере метаданных, в то время как сам файл хранится в объектном хранилище. Такой подход обладает рядом очевидных преимуществ при больших систем хранения,

среди которых практически неограниченные возможности масштабирования емкости хранилища, увеличение скорости доступа к данным и т.п.

2.7. Отказоустойчивость

Обеспечение надлежащего уровня отказоустойчивости является одним из ключевых аспектов организации современной системы хранения и, конечно, краеугольным камнем организации системы хранения экса-класса. За последние пять лет в этой области был достигнут значительный прогресс, некоторые решения оказались удивительно успешными, хотя в целом проблема обеспечения отказоустойчивости для системы экса-класса не решена [**26**]

Основная проблема в этой области, которая непосредственно затрагивает системы хранения связана с тем, что с увеличением масштабов вычислительных систем и приложений время сохранения контрольных точек становится значительным (15-30 минут) и приближается к MTBF.

Конечно, технологии увеличивающие пропускную способность систем хранения позволяют увеличить этот интервал, но основные надежды возлагаются на изменение технологий сохранения контрольных точек.

Одним из подходов является сокращение размеров контрольных точек, что требует ручной настройки приложений - программист должен указать наиболее критические участки и данные, которые следует сохранять.

Другой перспективный подход заключается в использовании многоуровневых технологий сохранения контрольных точек, что включает техники частичного перезапуска, сохранение чассти данных в памяти. Кроме того, данный подход включает использование комбинированных технологий хранения для оптимизации накладных расходов и надежности системы. В частности, в этом контексте упоминается наличие дисков хранения на вычислительных узлах ии/или использование burst buffera.

Другим очевидным подходом является репликация, которая, впрочем, связана с огромными накладными расходами на систему хранения.

2.8. RAID

Еще одной проблемой, связанной с организацией надежного хранилища, является явное устаревание технологии RAID, которая в настоящее время является наиболее узким местом в производительности дисковых массивов. Допустим, если возникла проблема с одним из дисков 4 или 6 TB в массиве, защищенном технологией RAID5 или RAID6, восстановление потребует огромного количества времени, в течение которого будет недоступен весь массив. В некоторых системах, недоступность одного массива приводит к отсутствию работоспособности всей файловой системы. Такая ситуация может возникнуть даже в случае, если поврежден всего один файл.

Интересное решение предлагает компания PANASAS [28], которое заключается в троекратном копировании данных, что увеличивает надежность в 150 раз. Для организации RAID6+, так названа эта технология, требуется 25% дополнительной емкости против 18% традиционного RAID6.

Кроме того, частью новой версии кластерной файловой системы PANASAS (PanFS6.0) является распределенный пофайловый RAID. Эта технология позволяет существенно сократить время на восстановление, при этом важно отметить, что чем больше дисков в системе, тем менее существенным будет время на восстановление.

2.9. Киберинфраструктура

Стоит отметить еще один критический аспект развития науки в эпоху "Четвертой парадигмы воспроизводимость. Общеизвестно, что воспроизводимость является одной из основных характеристик научного знания. Под этим термином понимается возможность повторить методы и результаты научного исследования. Этот вопрос, относящийся к философии науки, неожиданно и непосредственно влияет на область разработки систем хранения. Очевидно, что если в одной организации удается получить научный результат путем анализа большого объема данных, научное сообщество должно иметь доступ к начальным данным и результатам исследования. Наиболее оптимальным решением является создание специализированной национальной киберинфраструктуры, которая позволяет обеспечить доступ к большим массивам специализированных данных всем заинтересованным группам специалистов. Такая система действительно способствует

получению нового научного знания в эпоху больших данных. В мировой практике существует несколько успешных примеров. Как уже было указано выше, лидирующие суперкопьютеры Японии являются частью HPCI. В США существует несколько специализированных научных сетей, ориентированных на различные области науки, например инициатива iPlant.

3. Заключение

Проведенный обзор демонстрирует все увеличивающуюся роль соответствующей системы хранения в современных высокопроизводительных установках. Как отметил главный архитектор компании Xyratex Торбен Петерсен, система хранения больше не гражданин второго класса в области высокопроизводительных вычислений [17]. Современные контракты на разработку крупнейших суперкомпьютеров в первую очередь включают детальные требования к системе хранения, которая должна отвечать самым передовым требованиям к объему и пропускной способности. Выше описаны революционные подходы к созданию систем хранения, которые разрабатываются в рамках создания систем Cori и Trinity, которые представляют собой наиболее перспективные проекты США в области HPC и рассматриваются как задел для создания компьютера экса-класса. Ярким примером направления в Европе служит контракт между компанием Bull и немецким центром DKRZ на создание суперкомпьютера, который вошел бы в первую пятерку текущего рейтинга. По доступным данным, система хранения этого компьютера должна предоставлять объем в 45 PB, т.е. стать одной из крупнейших в мире [30].

На основании проанализированных фактов можно составить краткий прогноз основных характеристик систем хранения, которые будут занимать лидирующие позиции в ближайшем будущем.

Вероятнее всего, на аппаратном уровне доминирующие системы хранения данных будут многоуровневыми, и будут включать как минимум три базовых уровня: уровень SSD дисков, который будет предоставлять начальную пропускную способность хранилища, уровень традиционных жестких дисков и архивное хранилище на магнитных накопителях. Возможны модификации с несколькими уровнями жестких дисков, а также SSD накопителей, например, первый и самый производительный из которых будет использоваться как Burst

buffer, а второй - представять начальный уровень внешнего хранилища. Добавление уровней может способствовать улучшению соотношения стоимость/объем. Возможен также уровень внутрисистемного хранилища, возможно в пользовательском пространстве, как начальный уровень для сохранения критичесеих данных, в частности, контрольных точек [31]

Можно предположить, что непременным атрибутом лидирующих систем хранения станет Burst buffer, представленный дисками SSD.

По численным характеристикам показатели будут стремиться к 40-50 PB емкости (на уровнях жестких дисков и SSD, библиотеки на магнитных накопителях будут предоставлять дополнительные возможности, для хранения отработанных данных, интерес к которым маловероятен) и пропускной способности в 1000 GB/s. На настоящий момент это топовые показатели, и анализ доступной информации по текущим разработкам показывает, что именно к эти характеристики фигурируют в ряде ТЗ на создание вычислительных установок, которые строятся сегодня. Тем не менее, согласно отчету [32] в ближайшие пять лет для достижения уровня "большой петафлоп который является промежуточной ступенью для разработки машин эксакласса, потребуются системы хранения с объемом более 100 PB, а в ближайшие 10 лет для суперкомпьютера с мощностью в эксафлоп, вероятно, потребуется система хранения с объемом в эксабайт данных. Отдельное внимание будет уделено обеспечению надежности и целостности данных на подобных масштабных установках. На фоне того факта, что технологии RAID5 и RAID6 окончательно выйдут из игры, верх возьмут различные гибкие стратегии реплицирования и зеркалирования. Возможной, но маловероятной стратегией, тем не менее, представляется технология использования модифицированных вариантов защиты данных по технологиям RAID, например RAID6+, предложенный компанией Panasas.

На уровне программного обеспечения наиболее вероятным вариантом развития событий выглядит широкое использование объектных распределенных файловых систем хранения, в частности, модифицированных версий Lustre. Чтобы обеспечить эффективную работу Burst buffera и прочих уровней системы хранения, будут продолжены работы по созданию и модификации средств, позволяющих управлять перемещением данных между слоями. Эти средства должны будут обеспечивать прозрачный доступ ко всему гетерогенному пространству хранения и предоставлять возможность настройки гибких

стратегий автоматического перемещения и реплицирования данных между слоями. Кроме того, специальные, тщательно теоретически обоснованные стратегии будут созданы для обеспечения эффективного, масштабируемого процесса создания и сохранения контрольных точек в подобных многоуровневых системах.

Для стран, широко использующих высокопроизводительные вычисления в различных исследовательских и правительственных учреждениях, эффективным и перспективным вариантом представляется объединение вычислительных систем, и, в особенности, систем хранения, в единую киберинфраструктуру, так Data-грид. Такой подход позволит обеспечить доступ к одним и тем же данным нескольким группам исследователей, что упростит проведение междисциплинарных исследований, обмен полученными знаниями и проверку на воспроизводимость полученных результатов. Для обеспечения эффективной совместной работы в описанном гетерогенном Data-гриде потребуется наличие высокоуровневых средств управления данными, которые будут поддерживать большинство используемых программных и аппаратных решений и позволят создать единое многопользовательское пространство данных с гибкими настройками доступа и безопасности.

Список литературы

[1] Jahanian. *The Transformative Impact of Computing and Communication in a Data-Driven World* // Extreme Science and Engineering Discovery Environment (XSEDE) conference, Unknown Month July 13. (English) ↑1

[2] Jonathan Shaw. *Why "Big Data" Is a Big Deal* // Harward magazine, Apr 2014 Mar (English). ↑2

[3] Christy Pettey. *Gartner Reveals Top Predictions for IT Organizations and Users for 2013 and Beyond.* — ORLANDO, Fla, October 24, 2012. ↑2

[4] Anthony J. G. Hey. The Fourth Paradigm: Data-Intensive Scientific Discovery / (T. a. T. Hey Stewart and Tolle, ed.) Redmond, Washington: Microsoft Research, 2009. — 284 p. ↑2

[5] Е. О. Тютляева, А. А. Московский. *Анализ основных тенденций в области хранения данных* // Информационные технологии и вычислительные системы, 2012. Т. **2**, с. 64-75. ↑3

[6] Ph. D. Christopher G. Willard Addison Snell. *HPC User Site Census: Storage* // InterSect360 Research. Sunnyvale, CA 94088, January 2014. ↑9

[7] Ph. D. Christopher G. Willard Addison Snell. *HPC User Site Census: Storage* // InterSect360 Research. Sunnyvale, CA 94088, 2013. ↑14

[8] Marc Stearman. *Design and Installation of Sequoia's 55PB Lustre+ZFS File System* // HPC User Forum. — Dearborn, September 19, 2012. ↑

[9] *Stampede User Guide* Texas University, Texas University, 22 august 2014, URL https://www.tacc.utexas.edu/user-services/user-guides/stampede-user-guide. ↑8

[10] N. Liu, J. Cope, C. Carothers, R. Ross, G. Grider, A. Crume, C. Maltzahn. *On the Role of Burst Buffers in Leadership-Class Storage System* // Mass Storage Systems and Technologies (MSST) IEEE: IEEE 28th Symposium, 2012, p. 1–11. ↑11

[11] J. Dongarra. *Impact of architecture and technology for extreme scale on software and algorithm design* // Workshop on Cross-cutting Technologies for Computing at the Exascale Department of Energy, February 2010. (English) ↑11

[12] J Shalf. *Exascale computing technology challenges* // HEC FSIO Workshop. — Arlington, 2010, August, 2010. (English) ↑11

[13] P. Carns, K. Harms, W. Allcock, C. Bacon, S. Lang, R. Latham, R. Ross. *Understanding and Improving Computational Science Storage Access Through Continuous Characterization* // Trans. Storage, October 2011. Vol. **7**, no. 3, p. 8:1–8:26, URL http://doi.acm.org/10.1145/2027066.2027068. ↑11

[14] *The High-Performance Storage System (HPSS)* // User Guide OAK Ridge National Laboratory, OAK Ridge National Laboratory, URL https://www.olcf.ornl.gov/kb_articles/the-high-performance-storage-system-hpss/. ↑8

[15] Blaise Barney. *Using the Sequoia and Vulcan BG/Q Systems* // Tutorial Lawrence Livermore National Laboratory, Lawrence Livermore National Laboratory, URL https://computing.llnl.gov/tutorials/bgq/#ParallelIO. ↑8

[16] W. E. Allcoc. *Parallel File Systems at HPC Centers: Usage, Experiences, and Recommendations* // User Forum on Data-Intensive Computing NERSC. — Oakland, CA, 18 June 2014., URL https://www.nersc.gov/assets/Uploads/W01-DataIntensiveComputingPanel.pdf. (English) ↑8

[17] Torben Kling Petersen. *HPC Storage: Current Status and Future Directions* // HPC Advisory Council Switzerland Conference HPC Advisory Council and the Swiss Supercomputing Centre. — Lugano, Switzerland, 1 April 2014., URL http://www.hpcadvisorycouncil.com/events/2014/swiss-workshop/presos/Day_2/8_Xyratex.pdf. (English) ↑8, 17

[18] *Intrepid*: Our IBM Blue Gene/P Argonne National Laboratory. — Argonne, IL: Argonne Leadership Computing Facility., URL https://www.alcf.anl.gov/intrepid. ↑11

[19] N. Hemsoth. *Burst Buffers Flash Exascale Potential* Tabor Communications. — San Diego, CA: HPCwire, 1 May 2014., URL http://www.hpcwire.com/2014/05/01/burst-buffers-flash-exascale-potential/. ↑11

[20] *Trinity*: NERSC-8 Use Case Scenarios: SAND 2013-2941 Unclassified, Unlimited Release NERSC. — Berkeley, C, June 11, 2013, c. 9. ↑12

[21] O. Tatebe, K. Hiraga, N. Soda. *Gfarm Grid File System* // New Generation Comput, 2010. Vol. **28**, no. 2, p. 257-275, URL http://dblp.uni-trier.de/db/journals/ngc/ngc28.html#TatebeHS10. ↑

[22] M. Wan, R. Moore, A. Rajaseka. *Integration of Cloud Storage with Data Grids* // Third International Conference on the Virtual Computing Initiative Research Triangle Park. — North Carolina, US, Unknown Month October 22. ↑

[23] W. Schroeder. *iRODS the Integrated Rule Oriented Data-management System* // UCSD BigData Inaugural Workshop: video UC San Diego, November 25, 2013. ↑

[24] *iRODS Official wiki*: Who uses iRODS?: wiki. ↑10

[25] M. Feldman. *The Big Data Challenge*: Intelligent Tiered Storage at Scale: White Paper InterSect360 Research. — Sunnyvale, CA, November, 2013, p. 10. ↑13

[26] F. Cappello, A. Geist, S. Kale, B. Kramer, M. Snir. *Toward Exascale Resilience: 2014 Update* // Supercomputing Frontiers and Innovations, 2014. Vol. **1**, p. 1-28, URL http://superfri.org/superfri/article/view/14/7. ↑15

[27] P. Braam. *Exascale File Systems*: Scalability in ClusterStor's Colibri System: slides OpenFabrics Alliance. — USA, 2010, p. 26. ↑

[28] *PanFS RAID 6+*: Intelligent RAID for the Future: Panasas Official Blog Panasas. — Sunnyvale, CA, 2014., URL http://www.panasas.com/products/panfs/PanFS_RAID. ↑9, 16

[29] B. Dufrasne, B. Barbosa, P. Cronauer, J. F. Lepine, S. Manthorpe. *IBM System Storage DS8000 Easy Tie*: An IBM Redpaper publication IBM, 22 Aug 2013, p. 172., URL http://www.redbooks.ibm.com/redpapers/pdfs/redp4667.pdf. ↑

[30] M. Bottinger. *DKRZ and BULL sign the contract for HLRE-3*: Press Release DKRZ & Bull, May 2014, p. 2., URL http://www.dkrz.de/pdfs/presse-und-artikel/2014_Mai_PM_Bull-DKRZ_ENG_v1.pdf. ↑17

[31] R. Ross, P. Carns, D. Goodell, K. Harms, K. Iskra, D. Kimpe, R. Latham, T. Peterka, R. Thakur, V. Vishwanath. *Trends in HPC I/O and File Systems*: slides Argonne National Laboratory, p. 32. ↑18

[32] *Report of the Task Force on High Performance Computing of the Secretary of Energy Advisory Board (SEAB)*, 10 Aug 2014. — 30 с. ↑18

Об авторах:

Екатерина Олеговна Тютляева

Инженер в ИПС им. А.К. Айламазяна РАН с 2006 г, образование высшее, квалификация: математик, системный программист, автор более 11 печатных научных работ. Область научных интересов: системы хранения данных, высокопроизводительные вычисления, обработка больших объемов данных.

e-mail: orditad@gmail.com

Михаил Михайлович Тютляев

Руководитель отдела технической поддержки в компании Интерин-технологии. Научные интересы включают проведение научно-исследовательского анализа современных трендов в области хранения информации.

e-mail: michail@interin.ru

Образец ссылки на эту публикацию:

Е. О. Тютляева, М. М. Тютляев. *Системы хранения данных лидирующих суперкомпьютеров* // Программные системы: теория и приложения: электрон. научн. журн. 2014. Т. ??, № ?, с. ??–??.

URL: http://psta.psiras.ru/read/

Ekaterina Tyutlyaeva, Mikhail Tyutlyaev. *Top HPC Storages.*

Abstract. This paper examines major trends in HPC storage implementation field for most powerful computer systems, It includes specification comparative analysis for these systems. The paper explores key software and hardware storage features and states main HPC storage challenges. The probabilistic prediction of future top systems capability is worked out. (*in Russian*).

Key Words and Phrases: Data storage, Big Data, Top-500, Burst buffer, tiered storage.