Параллельные файловые системы с объектной архитектурой

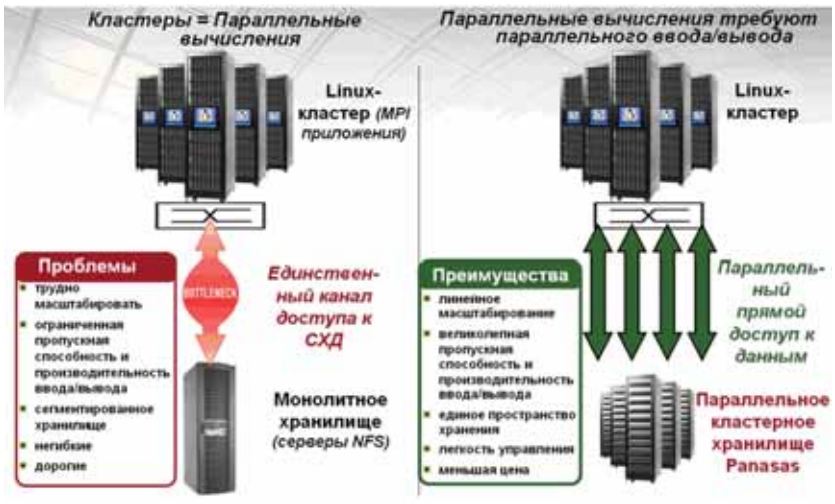
В последнее время в России наблюдается рост спроса на высокопроизводительные вычислительные системы в промышленности. Если сравнивать потребности промышленных предприятий в подобных решениях с вузами и научными институтами, то промышленности, прежде всего, необходима повышенная надежность и управляемость систем. Если говорить о собственно вычислительном ядре решения, то для кластерных систем оно само по себе является достаточно отказоустойчивым – выход какого-либо вычислительного узла из строя не влияет на общую работоспособность системы, а расчеты для большинства приложений могут быть продолжены с сохраненных контрольных точек. Однако все остальные компоненты системы должны быть продублированы или обладать повышенной надежностью. Так, с нашей точки зрения в промышленной высокопроизводительной вычислительной системе должны быть предусмотрены:

- система охлаждения и система электропитания с избыточностью;
- как минимум два управляющих узла, обеспечивающих работу пользователей;
- как минимум два выделенных узла, обеспечивающих мониторинг и управление системой;
- высокопроизводительная система хранения данных с обеспечением отказоустойчивости;
- система резервного копирования.

Особое внимание хотелось бы уделить вопросу выбора системы хранения данных. Практически все промышленные приложения работают с большими объемами данных и используют общую СХД для сохранения контрольных точек (их нельзя сохранять на локальных дисках вычислительных узлов, иначе выход узла из строя приведет к потере данных и расчеты придется начинать заново). Некоторые приложения также используют СХД для хранения рабочих данных в процессе счета. На рынке представлено достаточно много систем хранения, обладающих высокой производительностью и надежностью, но использование СХД для параллельных вычислений имеет одну специфическую особенность: все узлы кластерной системы должны иметь эффективный параллельный доступ к системе хранения данных. Это невозможно без использования специализированных параллельных файловых систем. Одним из наиболее критичных факторов здесь является масштабируемость – возможность сохранения высокой производительности при увеличении количества вычислительных узлов, работающих с хранилищем. В настоящее время существует лишь три файловые системы, которые могут эффективно использоваться на больших системах размером в сотни и тысячи узлов – это Lustre компании Cluster File Systems, PanFS компании Panasas и GPFS от IBM. Существуют и некоторые другие масштабируемые решения, но они не отвечают требованиям по обеспечению отказоустойчивости.

Чтобы обеспечить высокую масштабируемость решения, параллельные файловые системы используют объектную архитектуру*). Парадигма объектной архитектуры заключается в том, что каждый файл ассоциируется со множеством объектов, находящихся на разных физических устройствах. При выполнении файловых операций вычислительный узел получает "карту" с указанием расположения объектов на физических устройствах и далее имеет возможность напрямую работать с объектами, хранящимися на них, без использования выделенного сервера. Таким образом, обеспечивается множество параллельных путей доступа к данным с вычислительных узлов, что и позволяет добиться великолепной масштабируемости системы хранения. Из рассматриваемых файловых систем для промышленного применения наиболее интересно решение от компании Panasas. Lustre и GPFS, как правило, используют в качестве устройств хранения традиционные СХД с архитектурой SAN, подключенных к некоторому множеству серверов, являющихся серверами объектов данных. В отличие от них, Panasas предлагает законченное программно-аппаратное решение с блейд-архитектурой – своего рода кластер хранения данных. Такой подход обеспечивает непревзойденную простоту инсталляции и использования решения. В отличие от Lustre и GPFS, установка и настройка которых требует значительных усилий со стороны системного администратора, СХД от Panasas требует менее часа для подготовки к работе и минимального обслуживание в дальнейшем. Используя подключение через стандартную сеть Gigabit Ethernet, эта система может быть использована практически с любыми кластерными решениями, а расширение системы происходит путем простого подключения дополнительного модуля с блейдами, при этом объем файловой системы автоматически увеличивается прозрачно для пользователя.

*) из упомянутых файловых систем только Lustre и PanFS имеют полностью объектную архитектуру, GPFS использует комбинированную.



Все файлы в PanFS находятся в едином глобальном пространстве имен, что избавляет от необходимости миграции файлов на другую СХД. В то же время для разграничения доступа и назначения каких-либо конкретных свойств определенным группам файлов используется механизм виртуальных логических томов на уровне файловой системы. Важным свойством СХД от Panasas является также поддержка стандартных протоколов NFS и CIFS для доступа к данным. Таким образом, эта система может использоваться в качестве централизованного хранилища для гетерогенной инфраструктуры предприятия, что, например, позволяет избежать накладных расходов, связанных с переносом данных с рабочих станций, используемых для пре- и постпроцессинга, на вычислительную систему. Еще одной отличительной особенностью решения от Panasas является организация RAID на уровне файловой системы, а не физических устройств хранения. Это позволяет в рамках единого пространства имен реализовывать гибкую политику использования разных уровней RAID для определенных групп файлов. Система имеет богатый комплект встроенных средств диагностики, обеспечивающих как постоянный мониторинг состояния физических носителей с превентивной миграцией данных в случае обнаружения потенциальных проблем, так и мониторинг загрузки с динамической балансировкой в фоновом режиме.

Система также обладает высокой отказоустойчивостью – дублированные блоки питания, встроенная аккумуляторная батарея, обеспечивающая сброс дискового кэша и аккуратное завершение работы системы в случае полного отключения питания, возможность установки второго встроенного коммутатора Gigabit Ethernet для резервирования коммуникаций обеспечивают все необходимое для безотказной работы. Компания "Т-Платформы" предлагает параллельное кластерное хранилище T-Platforms® ReadyStorage ActiveScale Cluster, основанное на технологиях компании Panasas, а также богатый опыт конфигурирования, настройки и обслуживания таких хранилищ.

В заключение хотелось бы сказать несколько слов о перспективах развития суперкомпьютерной отрасли применительно к промышленным вычислительным системам. В последнее время очень большое внимание уделяется развитию альтернативных методов повышения производительности – использованию новых процессорных архитектур, различных аппаратных акселераторов, графических процессоров и т.п. На данный момент такие решения пока не очень распространены и не поддерживаются производителями коммерческого ПО. Дело в том, что модификация пользовательских программ для использования на альтернативных архитектурах является достаточно трудоемким процессом, поэтому использование таких специализированных решений в промышленных вычислительных системах пока нецелесообразно. Однако в ближайшие годы ситуация может существенно измениться: для более точных расчетов требуется существенно большая производительность систем, которую тяжело обеспечить с применением традиционных кластерных решений.

Андрей Слепухин,
руководитель Центра Кластерных Технологий, "Т-Платформы"