КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 543–547

Ки&М

СЕКЦИОННЫЕ ДОКЛАДЫ

УДК 004.4, 004.63

GridFTP frontend with redirection for DMlite

A. K. Kiryanov

Petersburg Nuclear Physics Institute, Orlova Roscha, Gatchina, 188300, Russia

E-mail: globus@pnpi.nw.ru

Received October 10, 2014

One of the most widely used storage solutions in WLCG is a Disk Pool Manager (DPM) developed and supported by SDC/ID group at CERN. Recently DPM went through a massive overhaul to address scalability and extensibility issues of the old code.

New system was called DMLite. Unlike the old DPM that was based on daemons, DMLite is arranged as a library that can be loaded directly by an application. This approach greatly improves performance and transaction rate by avoiding unnecessary inter-process communication via network as well as threading bottlenecks.

DMLite has a modular architecture with its core library providing only the very basic functionality. Backends (storage engines) and frontends (data access protocols) are implemented as plug-in modules. Doubtlessly DMLite wouldn't be able to completely replace DPM without GridFTP as it is used for most of the data transfers in WLCG.

In DPM GridFTP support was implemented in a Data Storage Interface (DSI) module for Globus' GridFTP server. In DMLite an effort was made to rewrite a GridFTP module from scratch in order to take advantage of new DMLite features and also implement new functionality. The most important improvement over the old version is a redirection capability.

With old GridFTP frontend a client needed to contact SRM on the head node in order to obtain a transfer URL (TURL) before reading or writing a file. With new GridFTP frontend this is no longer necessary: a client may connect directly to the GridFTP server on the head node and perform file I/O using only logical file names (LFNs). Data channel is then automatically redirected to a proper disk node.

This renders the most often used part of SRM unnecessary, simplifies file access and improves performance. It also makes DMLite a more appealing choice for non-LHC VOs that were never much interested in SRM.

With new GridFTP frontend it's also possible to access data on various DMLite-supported backends like HDFS, S3 and legacy DPM.

Keywords: WLCG, Grid, GridFTP, DPM, DMLite, data storage, access protocol

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 3, pp. 543–547.

**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 543−547**

Ки&М

СЕКЦИОННЫЕ ДОКЛАДЫ

Поддержка протокола GridFTP с возможностью перенаправления соединений в DMLite Title

А. К. Кирьянов

Петербургский институт ядерной физики им., Россия, 188300, Ленинградская обл., Гатчина, Орлова роща, ФГБУ ПИЯФ

Одним из наиболее широко используемых решений для хранения данных в WLCG является Disk Pool Manager (DPM), разрабатываемый и поддерживаемый группой SDC/ID в ЦЕРНе. Недавно старый код DPM был практически переписан с нуля для решения накопившихся проблем с масштабируемостью и расширением функциональности.

Новая система была названа DMLite. В отличие от DPM, который был реализован в виде нескольких демонов, DMLite выполнена в виде программной библиотеки, которая может быть непосредственно загружена приложением. Такой подход значительно повышает общую производительность и скорость обработки транзакций, избегая ненужного межпроцессного взаимодействия через сеть, а также узких мест в многопоточной обработке.

DMLite имеет модульную архитектуру, при которой основная библиотека обеспечивает только несколько базовых функций. Системы хранения данных, а также протоколы доступа к ним реализованы в виде подключаемых модулей (plug-ins). Конечно, DMLite не смогла бы полностью заменить DPM без поддержки протокола GridFTP, наиболее широко используемого для передачи данных в WLCG.

В DPM поддержка протокола GridFTP была реализована в виде модуля Data Storage Interface (DSI) для GridFTP сервера Globus. В DMLite было решено переписать модуль GridFTP с нуля, чтобы, во-первых, воспользоваться новыми возможностями DMLite, а во-вторых, добавить недостающую функциональность. Наиболее важным отличием по сравнению со старой версией является возможность перенаправления соединений.

При использовании старого интерфейса GridFTP клиенту было необходимо предварительно связаться со службой SRM на головном узле хранилища, чтобы получить Transfer URL (TURL), необходимый для обращения к файлу. С новым интерфейсом GridFTP делать этот промежуточный шаг не требуется: клиент может сразу подключиться к службе GridFTP на головном узле хранилища и выполнять чтение-запись используя логические имена файлов (LFNs). Канал передачи данных при этом будет автоматически перенаправлен на соответствующий дисковый узел.

Такая схема работы делает одну из наиболее часто используемых функций SRM ненужной, упрощает доступ к файлам и повышает производительность. Это также делает DMLite более привлекательным выбором для виртуальных организаций, не относящихся к БАК, поскольку они никогда не были особо заинтересованы в SRM.

Новый интерфейс GridFTP также открывает возможности для хранения данных на множестве альтернативных систем, поддерживаемых DMLite, таких как HDFS, S3 и существующие пулы DPM.

Ключевые слова: БАК, Грид, хранилище данных, протокол доступа

One of the most widely used storage solutions in WLCG is a Disk Pool Manager [DPM..., 2015] developed and supported by SDC/ID group at CERN. It was started in 2005 and over the time built a strong install base on more than 200 sites. Unfortunately some of the architectural decisions taken at the very beginning of the project resulted in scalability and extensibility issues such as:

– Monolithic code base, that was hard to maintain and add features to.
– IPC between multiple daemons even for simple operations.
– Reliance on SRM which turned out not to be attractive outside of HEP community.

Eventually it was decided to do a massive overhaul of the old code and effectively build a completely new system that would be free of DPM shortcomings while maintaining backwards compatibility with old clients: DMLite [DMLite..., 2015]. Unlike the old DPM that was based on daemons, DMLite is arranged as a library that can be loaded directly by an application. This approach greatly improves performance and transaction rate by avoiding unnecessary inter-process communication via network as well as threading bottlenecks.

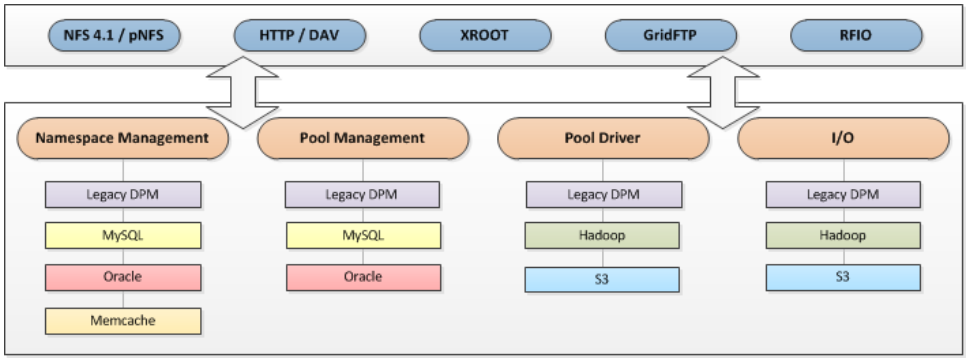

Fig. 1. DMLite modular architecture

DMLite has a modular architecture (fig.1) with its core library providing only the very basic functionality. Backends (storage engines — on bottom) and frontends (data access protocols — on top) are implemented as plug-in modules. While HTTP and XROOT are gaining momentum in inter-site data transfers, doubtlessly DMLite wouldn't be able to completely replace DPM without GridFTP [GridFTP, 2015] as it is still used for most of the data transfers in WLCG.

In DPM GridFTP support was implemented in a Data Storage Interface (DSI) module for Globus Toolkit GridFTP server. In DMLite an effort was made to rewrite GridFTP module from scratch in order to take advantage of new DMLite features and also implement new functionality. The most important improvement over the old version is redirection: an ability to authenticate clients and accept data transfer requests at one point (host) but seamlessly perform an actual data exchange with another one.

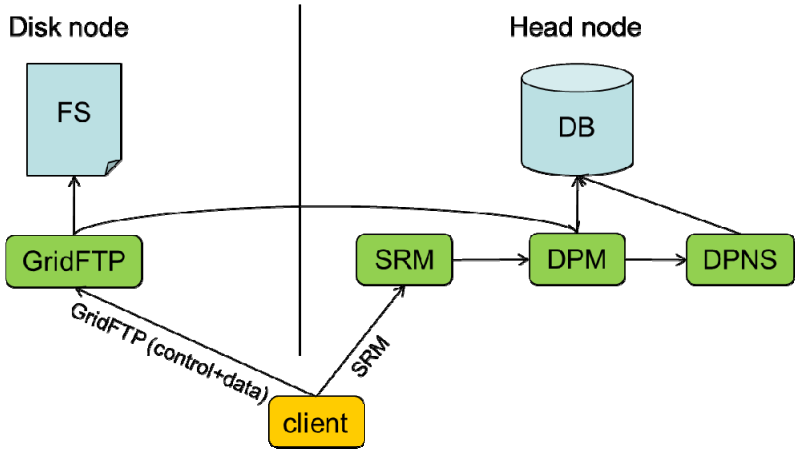

Fig. 2. GridFTP access with DPM

With old DPM GridFTP frontend a client needed to contact SRM on the head node in order to convert logical filename (LFN) to a transfer URL (TURL) before reading or writing a file (fig. 2). With new GridFTP frontend this is no longer necessary: a client may connect directly to the GridFTP server on the head node and perform file I/O using only logical file names (LFNs). Data channel is then automatically redirected to a proper disk node (fig. 3). This renders the most often used part of SRM unnecessary, simplifies file access and improves performance. It also makes DMLite a more appealing choice for non-LHC VOs that were never much interested in SRM.

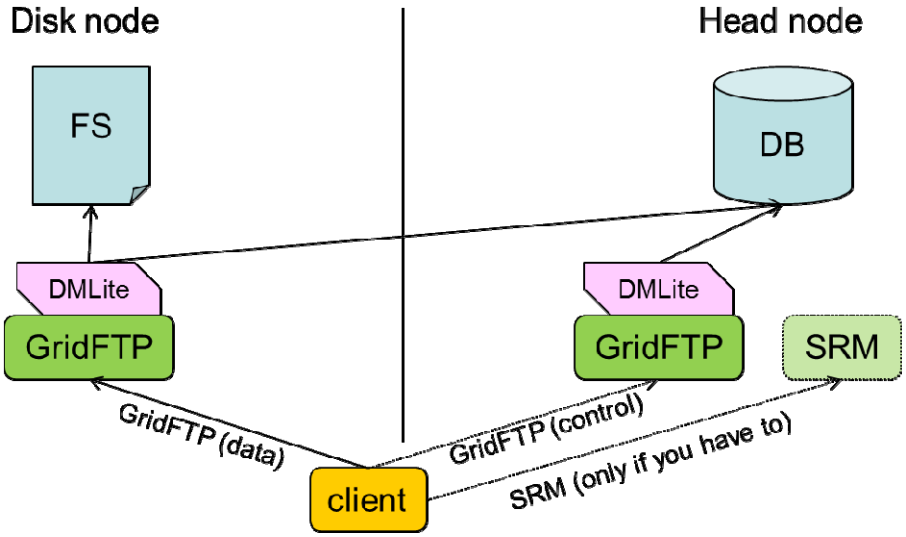

Fig. 3. GridFTP access with DMLite

GridFTP redirection was implemented thanks to the Delayed Passive connection mode available as one of the GridFTP v2 extensions available in Globus Toolkit. With legacy Passive mode a server had to provide connection endpoint parameters (address and port) before a file request, at a time point when file name was not yet known. With Delayed Passive a server postpones its response until a client actually initiates a file transfer, which makes it possible to redirect a client to different disk nodes depending on file name (fig. 4).

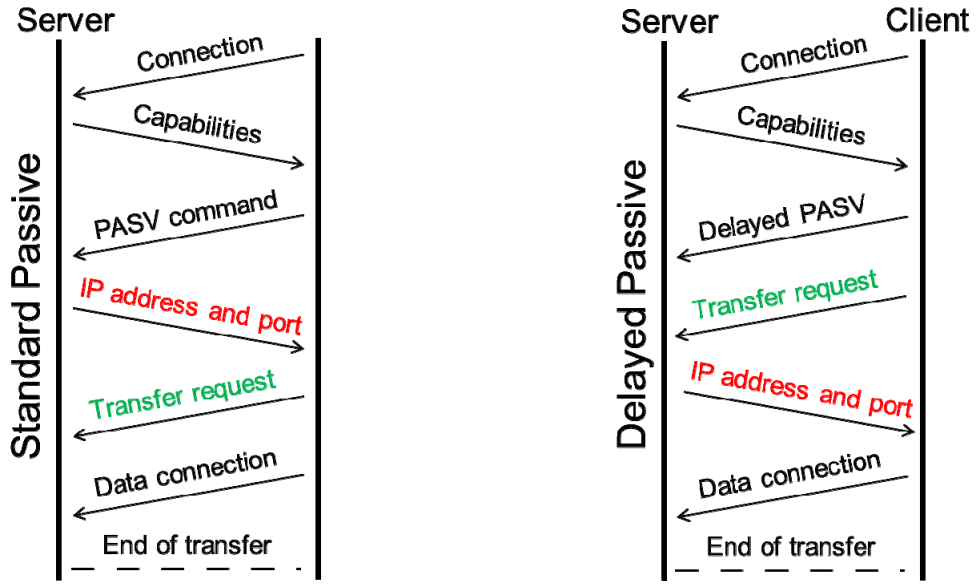

Fig. 4. Passive mode vs. Delayed Passive mode

New GridFTP frontend supports two deployment scenarios: stand-alone and redirecting. Stand-alone scenario is essentially identical to the way GridFTP was deployed with old DPM: GridFTP servers are configured independently and a client has to know which server to connect to. This may be used as drop-in replacement for DPM+SRM installations.

Redirecting scenario is somewhat different: head node is configured as a redirector, and disk nodes are put in “backend mode” which forbids direct client connections. By this scenario a client always has to contact head node and does not have to worry about obtaining TURLs to a proper disk node. The only limitation of this scenario is that clients have to support Delayed Passive mode for optimal performance.

GridFTP frontend for DMLite is in production since mid-2014. It is supported by GFAL2 library and FTS3 file transfer service which is widely used on WLCG for scheduled file transfers. With new GridFTP frontend it's also possible to access data on various DMLite-supported backends like HDFS, S3 and legacy DPM.

References

DPM (Disk Pool Manager) [online] // — 2015 — URL: https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm/Dev/Dmlite (дата обращения: 17.01.2015);

DMLite [online] // — 2015 — URL: https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm/Dev/Dmlite (дата обращения: 17.01.2015);

GridFTP [online] // — 2015 — URL: http://home.web.cern.ch/about/computing/worldwide-lhc-computing-grid (дата обращения: 17.01.2015).