SFS INTEGRÁLÁSA GRID RENDSZERHEZ

Dóbé Péter, dobe@iit.bme.hu
Dr. Szeberényi Imre, szebi@iit.bme.hu
BME-IIT – BME-IK

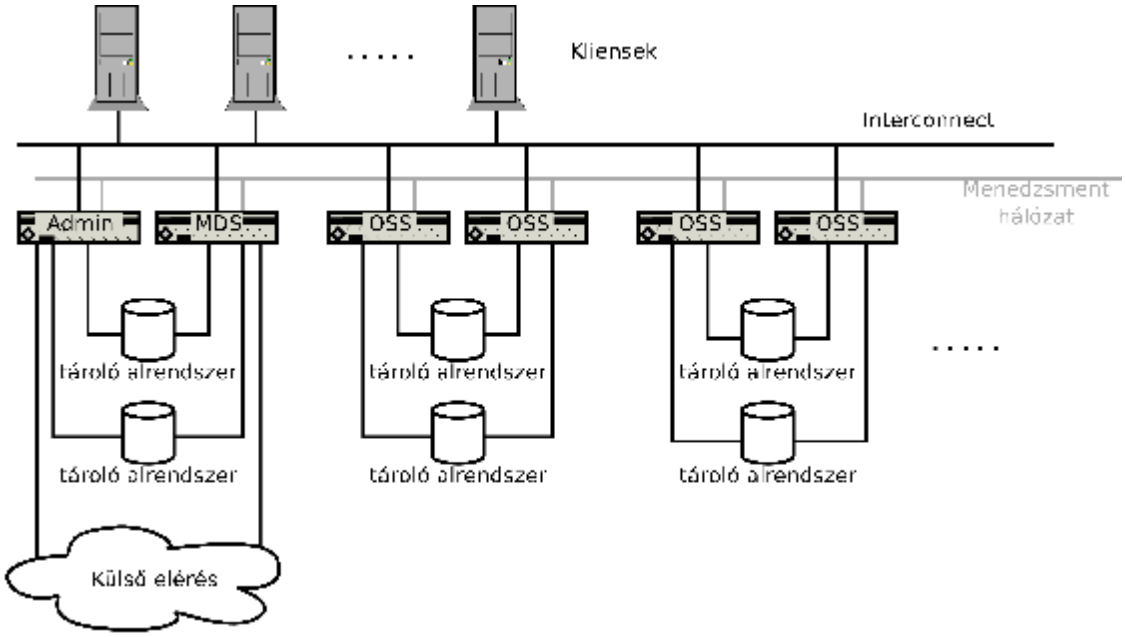
A BME Informatikai Központban számos témában folyik kutatás-fejlesztési tevékenység, többek között a Griddel és párhuzamos számítással kapcsolatban is [1]. Ennek során a központ szorosan együttműködik ipari partnerekkel. Ilyen együttműködés keretében született az itt bemutatott eredmény is, ahol az európai Gridet kibővítettük nagy méretű és hatékonyan elérhető tárkapacitással.

1. HP Scalable File Share általános jellemzői

A Scalable File Share (SFS) a HP által gyártott nagy kapacitású adattároló rendszer [2], amely a nyílt forráskódú Lustre állományrendszeren alapul [3]. Ez a kifejezetten klaszteres alkalmazásra, párhuzamos számítási feladatok támogatására kifejlesztett POSIX-kompatibilis állományrendszer rendkívül jól skálázható: akár több mint tízezer klienst kiszolgálhat és petabájt nagyságrendű adatot tárolhatunk rajta.

Az SFS független szervergépekből és a hozzájuk kapcsolódó, több lemezegységből álló, RAID szervezésű tároló alrendszerekből épül fel. (ld. 1. ábra). A rendszer kezelése az Adminisztrációs Szerveren keresztül történik. A tárolni kívánt adat és a metaadat külön tároló alrendszeren kapott helyet, és külön szerver szolgálja ki. Az előbbieket kiszolgáló szerver neve Meta Data Server (MDS), az utóbbiakért az Object Storage Server-ek (OSS) felelnek. Az OSS-ek és a hozzájuk tartozó tároló alrendszerek számának növelésével a rendszer könnyen skálázható.

A szerverek párokra vannak osztva, ahol az Adminisztrációs Szerver és az MDS egy párba tartozik. Amennyiben az MDS üzemképtelenné válik, az Adminisztrációs Szerver átveszi a szerepét. Hasonlóan, ha az Adminisztrációs Szerver esik ki, az MDS helyettesítheti. Az egy párban levő OSS-ek osztoznak a tároló alrendszereken, így az egyik szerver hibája esetén is elérhető a rajtuk található adat. Ez a kialakítás nagy mértékű hibatűrést eredményez.



1. ábra: SFS rendszer felépítése

Három különböző hálózati kapcsolat köti össze az SFS-t. A szervereket egy menedzsment hálózat köti össze, amelynek nincs külső kapcsolata. Az Adminisztrációs Szerver és az MDS kívülről elérhető, ezen a kapcsolaton keresztül lehet menedzselni a szervereket. Az adatkapcsolat a kliensekkel egy harmadik hálózaton (Interconnect) keresztül folyik. Ez utóbbi kapja a legnagyobb terhelést, hiszen itt történik a tárolt állományok elérése, amelyet egyszerre sok kliens is végezhet, ezért ennek nagy sebességűnek kell lenni. A legegyszerűbb konfigurációban ez egy Gigabit Ethernet hálózat, amellyel 110 MB/s adatsebességet lehet elérni, azonban használhatók más technológiák Interconnect célra, például Quadrics, Myrinet [4] vagy InfiniBand [5] is, így akár 770 MB/s sebesség is elérhető.

2. Az EGEE és a gLite köztesréteg

Az Enabling Grids for E-sciencE (EGEE) [6] projekt az Európai Unió által támogatott legnagyobb Grid projekt, amelyben 32 ország vesz részt. Ez a Grid közel 30 ezer processzorból áll, továbbá mintegy 5 petabájt adatot képes tárolni. Ezt a hatalmas mennyiségű erőforráskészletet általános tudományos célra felhasználhatják azok a kutatók, akik egy virtuális szervezetnek (Virtual Organization, VO) tagjai. Fejlesztése a svájci CERN kutatóközpontban folyik; eredetileg az LHC (Large Hadron Collider) részecskegyorsítóból érkező adatokat kiszolgáló Gridből, az LCG-ből (LHC Computing Grid) indult ki.

Az EGEE használatához szükséges szoftver a gLite [7] köztesréteg, amely jelenleg a Scientific Linux operációs rendszerre telepíthető. A Grid elérése egy olyan számítógépen történhet, amely rendelkezik a szükséges parancssori vagy grafikus felhasználói felülettel (UI, User Interface). Az UI számítógép a WMS-hez (Workload Management System) fordul, melynek feladata a végrehajtandó feladat eljuttatása a futtató erőforrásokhoz. Egy-egy erőforráscsoportot egy CE (Computing Element) fog össze, ennek központi számítógépe a GG (Grid Gate). Az adattárolásért az SE (Storage Element) gép felel, amelyen keresztül gridFTP-vel és egyéb protokollokkal hozzáférhetők az adatok. A CE-hez tartozik egy helyi ütemező, amely kiosztja a feladatokat munkacsomópontoknak (Worker Node, WN), továbbá információs szolgáltatások, amelyek az erőforrás-felderítésben vesznek részt.

Az erőforrásokat igénybevevő felhasználó hitelesítése aszimmetrikus kulcsú titkosításon és X.509 tanúsítványokon alapul, így a felhasználó a tanúsítványa felmutatásával igazolja, hogy az adott VO tagja, és a titkos kulcsa segítségével használatba tudja venni a szolgáltatásokat. A VO tagságok kezelése a VOMS (Virtual Organization Membership Service) használata révén lehetséges.

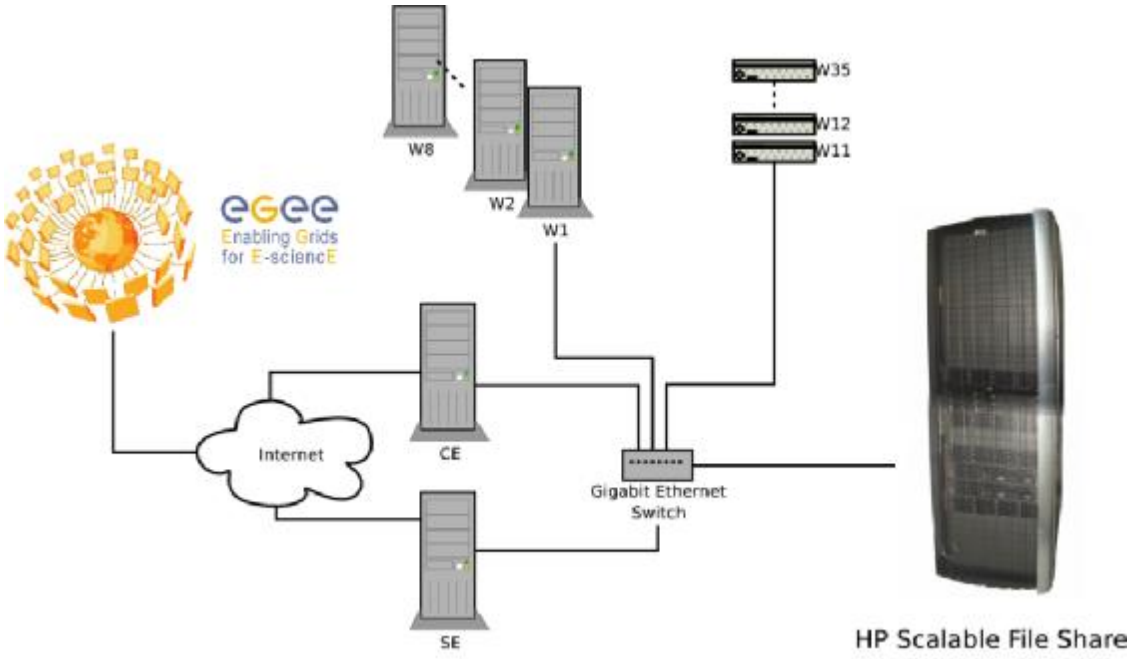
3. Integrálás az EGEE-hez

A Budapesti Műszaki és Gazdaságtudományi Egyetem Informatika Központjában rendelkezésünkre áll egy HP SFS rendszer kísérleti célokra. Ez jelenleg üzemel, közel 3 terabájt adatot képes tárolni. Négy szerver található benne: az Adminisztrációs Szerver, az MDS és két OSS, mindegyikhez SFS20 típusú tároló alrendszerek kapcsolódnak. Az Interconnect hálózat Gigabit Ethernet, amely viszonylag kis sebességű, ezért az adatátvitelnél ez jelenti a szűk keresztmetszetet.

A BME-n szintén összeállítottunk egy számítási erőforráskészletet, amely az EGEE-nek részét képezi (ld. 2. ábra). Ennek központi rendszerén, amely a CE nevet kapta, fut többek között a GG és a WMS szolgáltatás, emellett egy dedikált SE szervergép is működik. E két számítógép, valamint további nyolc, melyek WN-ként üzemelnek, HP ProLiant G2 szerverek két Intel Xeon processzorral és helyi SATA diszkekkel. A diszkek egyrészt a beküldött feladatok ideiglenes tárhelyeként szolgálnak, másrészt a Grid számára adattárolási erőforrásként is használhatók. További számos WN gép van, ezek kisebb teljesítményű asztali PC-k.

A HP szerverek egyező hardver konfigurációja lehetővé teszi, hogy ne kelljen mindegyiken a szoftvert külön beállítani, hiszen ugyanaz az operációs rendszer kernel futhat mindegyiken. Ezért azt a megoldást választottuk, hogy az azonos funkciójú számítógépek a szükséges szoftver és adatok nagy részét nem a helyi diszkeken tárolják, hanem egy központi NFS szerveren érik el. Ezáltal nemcsak a konfigurációval eltöltendő időt csökkentjük le, hanem lemezterületet is megtakarítunk. Így ugyanis azok mellett az adatok mellett, amelyeket mindenképpen a helyi lemezen kell tárolni (ilyen például a hitelesítéshez szükséges egyedi titkos kulcs), több tárterület marad a futtatandó feladatok adatai számára.

Az összes résztvevő szervergép az egyetemi SFS Interconnect hálózatára is rákapcsolódik, ezért kliensként használni tudják azt. A WN-ek tehát egy rendkívül nagy kapacitású, gyors elérésű stabil tárterületet használhatnak, ugyanakkor ezt az állományrendszert az SE gépen keresztül a Grid felhasználói is igénybevehetik általános adattárolásra.



2. ábra: EGEE erőforráskészlet SFS-sel

A többprocesszoros szervergépek nyújtotta számítási teljesítmény és a 3 terabájt tárkapacitás ezáltal számos virtuális szervezet rendelkezésére áll, amelyek kutatásra használják a Gridet: ilyen például az Atlas [8] és az Alice [9] VO. Ezen kívül létrehoztunk egy saját virtuális szervezetet, „egeebme” néven. Ez egyelőre kísérleti illetve oktatási célt szolgál, tagjai a BME hallgatói és kutatói.

4. Köszönetnyilvánítás

E munka részben a Nemzeti Kutatási és Technológiai Hivatal Pázmány Péter programjának (RET-06/2005) támogatásával jött létre. A szerzők szeretnék kifejezni a köszönetüket az Európai Unió által támogatott EGEE projektnek (EU INFSO-RI-031688), valamint az NKFP MEGA (2_009_04) projektnek.

Hivatkozások:

[1] I. Foster, C. Kesselman: *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann Publishers, 1998

[2] *HP StorageWorks Scalable File Share*, http://www.hp.com/techservers/products/sfs.html

[3] *Lustre*, Cluster File Systems Inc. http://www.lustre.org/

[4] *Myrinet*, ANSI/VITA 26-1998 http://www.myricom.com/myrinet/overview/

[5] *InfiniBand Trade Association*, http://www.infinibandta.org/

[6] *Enabling Grids for E-sciencE*, http://www.eu-egee.org/

[7] *gLite: Lightweight Middleware for Grid Computing*, http://glite.web.cern.ch/glite/

[8] *The ATLAS experiment*, http://atlasexperiment.org/

[9] *ALICE: A Large Ion Collider Experiment at CERN LHC*, http://aliceinfo.cern.ch/index.html