ADATBÁNYÁSZ ALKALMAZÁS TÁMOGATÁSA A CLUSTERGRIDBEN

Vida Gábor, vida@sztaki.hu
Dr. Podhorszki Norbert, pnorbert@sztaki.hu
Dr. Kacsuk Péter, kacsuk@sztaki.hu
MTA SZTAKI Laboratory of Parallel and Distributed Systems

Tematika:

A ‘’ Következő generációs adatbányászat nagy teljesítményű elosztott párhuzamos rendszereken” c. GVOP projekt fő célja egy nagy teljesítményű, elosztott, párhuzamos rendszereken működő adatbányászati szoftver prototípus kifejlesztése, lehetővé téve minden korábbinál jobb minőségű adatbányászati modellek létrehozását. A rendszer feladata, hogy a hálózati rendszer tulajdonságaihoz alkalmazkodva a rendelkezésére álló erőforrásokat a lehető legoptimálisabban használja fel, miközben az adatbányászattal foglalkozó szakembert mentesíti a párhuzamos feldolgozással kapcsolatos részletkérdések ismeretétől, miközben folyamatos tájékoztatást kap a részeredményektől, s ezek függvényében beavatkozhat a folyamatokba. A projekt keretében a nagy adat- és számításigényes adatbányász alkalmazásokat Grid rendszerek segítségével hajtjuk végre.

Az elosztott számítási környezet, konkrét Gridek elfedésére és a megfelelő funkciók biztosítására egy új programozói felületet definiáltunk a projektben (Distributed Computing API, vagy röviden DC API). A mögöttes implementáció feladata az elosztott rendszer elérésével kapcsolatos feladatok ellátása, a működéshez szükséges hálózatkezelés, illetve az adatszolgáltatás megvalósítása. A rá épülő rétegek igényeit kiszolgálja, de elfedi az elosztott párhuzamos rendszer implementációjának részleteit. A DC API alapimplementációjának biztosítania kell azoknak az alapfunkciónak a működését, amelyek az első futtatható demo verzió futtatását lehetővé teszik.

A DC API első implementációját a ClusterGriden végeztük el és a jelen cikk célja ennek az implementációnak a bemutatása. A cikkben bemutatjuk a munkacsomagok létrehozásának és Gridben történő kiosztásának módját, a munkacsomagok végrehajtásának, felfüggesztésének, megszakításának és újraindításának technikáját. Az implementáció során a legnagyobb kihívást a részeredmények kezelése és ennek alapján a végrehajtás on-line vezérlésének megvalósítása okozta, ezért ennek megoldását részletesebben is taglalja a cikk.