Računalni klasteri

Sadržaj

Paralelno računarstvo

U današnje vrijeme teško je zamisliti život bez računala, koji su kroz godine postajali sve “jači” i “brži”. Budući da se kroz godine razvoja računala došlo do fizičkih limita tehnologije poluvodiča sve je teže pronaći nove načine za ubrzavanje procesa i zadovoljavanje sve većih potreba naprednijih korisnika. Mnoge grane znanosti i ljudske djelatnosti (fizika, farmacija, inžinjerski problemi, ekonomija, vojne primjene...) imaju potrebu za ogromnom računarskom snagom koju ne mogu zadovoljiti ni najjače radne stanice. Iz tog razloga pribjegava se paralelizaciji procesa.

Paralelizacijom omogućavamo da se problem podjeli na više manjih radnji, operacija ili proračuna koji se izvode istovremeno. Imajući na umu da u današnje vrijeme i najslabija stolna računala imaju najmanje dvije CPU jezgre, možemo zaključiti da su sva današnja računala zapravo mali paralelni sustavi. Naravno, da bi paralelizacija uopće bila moguća, aplikacije moraju podržavati takav način rada, što razvoj aplikacija čini znatno složenijim nego prije. Ukoliko naši algoritmi nisu optimizirani za ovakav način rada, paralelizacijom procesa vrijeme izvođenja paralelne varijante biti će jednako ili čak i duže, nego u serijskoj varijanti.

Paralelno sklopovlje

Superračunala su računala koja u trenutku svog nastanka spadaju među najmoćnija dostupna računala za zahtjevne računalne probleme. Procesorska snaga superračunala često se koristi za rješavanje samo jednog specifičnog problema, te su takva superračunala obično hardverski i softverski prilagođena vrsti zadataka koje će morati obavljati.

Superračunala se danas dijele na tri osnovne vrste:

- vektorska superračunala - Cray SV, NEC SX
- velika paralelna računala (MPP – eng. Massively parallel processors) - IBM BlueGene/Q, Cray XC40
- računalni klasteri – sastoje se od više međusobno povezanih računala koja pomoću softwarea funkcioniraju poput jednog superračunala.

Računalni klasteri

Uopćeno gledajući, **klaster** je skupina nezavisno djelujućih elemenata povezanih nekim medijem u cilju koordiniranog i kooperativnog ponašanja. Računalni klaster nije precizno definiran pojam pa se čak i skup nezavisnih računala može smatrati klasterom ukoliko postoji neki minimalan vid i stupanj integracije.

Ipak, računalnim klasterom obično smatramo **sustav računala umreženih korištenjem brze lokalne mreže (LAN) pomoću koje računala međusobno komuniciraju**. Korištenje specifične programske podrške daje visok stupanj integracije računala, omogućava njihov koordinirani zajednički rad i pretvara ih efektivno (u mjeri u kojoj je to fizički moguće) u jedinstven višeprocesorski sustav. Neizbježna heterogenost arhitekture te korištenje distribuirane (umjesto podjeljene) memorije jedino je po čemu se klaster razlikuje od jedinstvenog višeprocesorskog sustava.

Ciljevi povezivanja računala u klastere su razni, a najčešće se računala povezuju s ciljem osiguravanja veće pouzdanosti ili većih performansi u odnosu na pojedino računalo.

Klastere možemo podijeliti na:

- klastere s visokom učinkovitošću – HPC – *eng.* High Performance Computing
- klastere s visokom propusnošću – HT – *eng.* High Throughput
- klastere s visokom dostupnošću – HA – *eng.* High Availability
- klastere za ravnomjerno opterećenje – LB – *eng.* Load Balancing
- hibridne klastere (sadrže specifične uređaje – matematičke uređaje, Graphics Processing Unit...)

Računalni klaster Isabella spada u HPC skupinu klastera, odnosno Beowulf HPC klastere. Računalni klasteri su široko rasprostranjeni u paralelnom računarstvu i čine više od 80% Top500 liste najmoćnijih računala.

BEOWULF

Beowulf je naziv za koncept HPC klastera koji je 1994. godinerazvijen u NASA Goddard Space Flight centru. Tada je korištenjem 16 komercijalno dobavljivih računala baziranih na Intel 100MHz 486 procesorima, povezanim dvostrukim 10 Mbps Ethernet LAN-om te pogonjenih Linux operativnim sustavom i PVM (Paralel Virtual Machine) bibliotekama izgrađen klaster kojim su demonstrirane primjenjivost, performanse i financijska isplativost Beowulf sustava u znanstvenim računalnim aplikacijama.

Ovakav sustav pokazao se u mnogim aplikacijama usporediv s dotadašnjim konceptima razvoja superračunala koja uglavnom koriste pristup u kojem se unutar istog fizičkog računala za obradu koristi više međusobno povezanih i paralelnih procesora (simetričnih višeprocesorskih sustava) koji međusobno dijele ostale resurse računala pa tako i komuniciraju korištenjem zajedničke **dijeljene memorije** (*shared memory*). Za razliku od ovog pristupa klaster koristi koncept labavog povezivanja računala na kojima se odvijaju procesi koji međusobno komuniciraju korištenjem raznih biblioteka za prijenos poruka te na taj način razvijaju koncept **distribuirane memorije** (*distributed memory*).

Da bi koncept distribuirane memorije dobro funkcionirao uobičajeno je da se za prijenos poruka, uz lokalnu mrežu koja se koristi za prijenos podataka sa zajedničkog podatkovnog sustava, koristi i posebno dodijeljena lokalna mreža kojom su povezana računala - **računalni nodovi** u klasteru. Ova mreža osim što omogućava prijenos velike količine podataka mora osiguravati i male latencije u prijenosu podataka (manje od 1 ms) pa se stoga koriste specifične izvedbe kao što su Myrinet ili MVIA.

Zahvaljujući Beowulf konceptu kojim je uz korištenje lako dobavljivih, odnosno komercijalnih i besplatnih komponenti (poput osobnih računala, Linux operativnog saustava i drugih) na jednostavan način moguće brzo izgraditi računalni klaster, u svijetu je ubrzo izgrađen velik broj sličnih klastera koji djeluju u znanstvenim i akademskim sredinama.

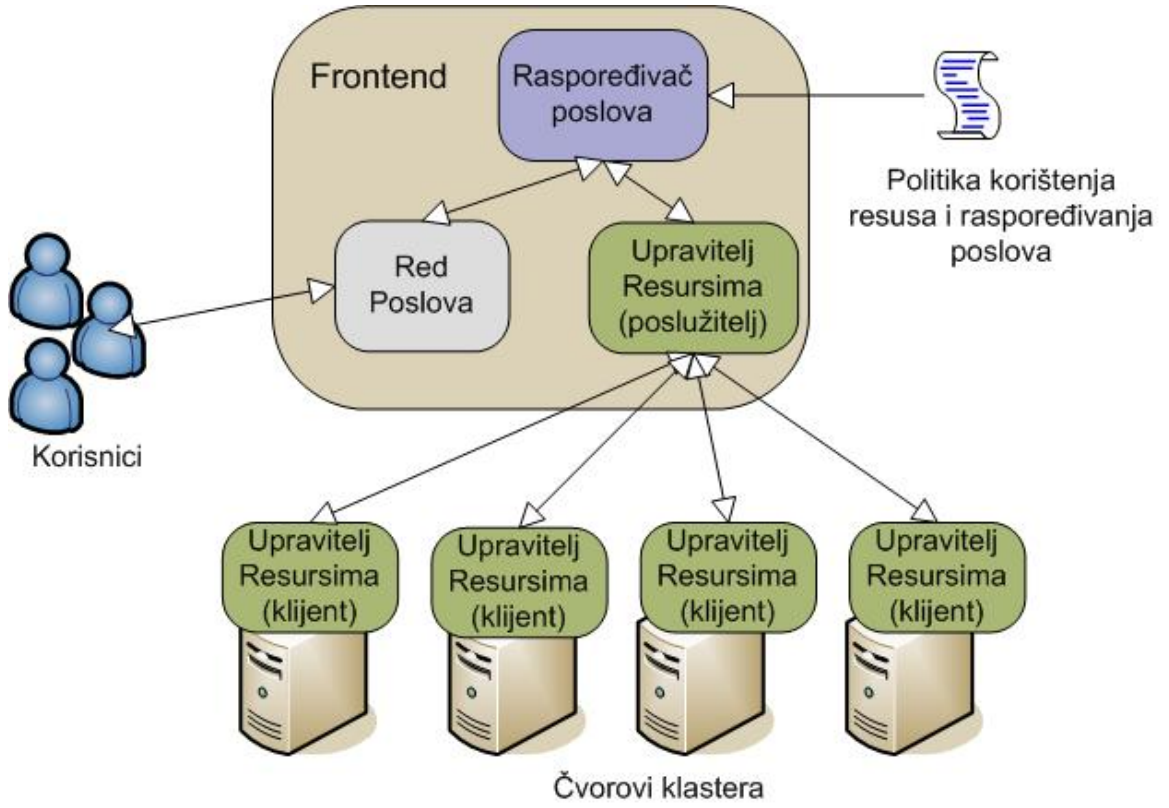
S obzirom da je besplatan, prilagodljiv i ima otvoreni kod, Linux je ubrzo postao najprihvatljiviji operativni sustav za izgradnju klastera. Na osnovi najraširenijih distribucija Linuxa – Red Hat i Debian izgrađeni su gotovi paketi za izgradnju klastera od kojih je jedan - **NPACI Rocks Clustering Toolkit** - korišten i u izvedbi računalnog klastera Isabella.

Sustavi za upravljanje poslovima

Sustavi za upravljanje poslovima (JMS) upravljaju izvođenjem korisničkih aplikacija na klasterima (vidi članak "Sustavi za upravljanje poslovima"). Slika prikazuje arhitektura tipičnog JMS-a. JMS-ovi su centralizirani sustavi čiji su središnje komponente smještene na zasebnom računalu. Kod Beowulf klastera je to *frontend*.

Sustav za upravljanje poslovima se sastoji tri komponente:

- Reda poslova (engl. *Queue Manager*)
- R*asporeñivača poslova* (engl. *Scheduler*)
- *Upravitelja resursima* (engl. *Resource Manager*).



Korisnik pomoću naredbi JMS-a (npr. qsub, condor_submit, itd.) pokreće aplikacije. Zahtjevi se spremaju u nizove poslova u kojima čekaju početak izvođenja. Primjer stanja u redu poslova na produkcijskom klasteru **Isabella** koji koristi sustav za upravljanje poslovima SGE:

Prva četiri posla su aktivna, a 5. čeka na izvođenje.

```
job-ID    name         user     state submit/start
-----------------------------------------------------------
3504    f2b14f2.ru   gkovacev     r     11/19/2004 01:41:01
3426    s2312-153.   Mpavicic     r     11/12/2004 22:35:22
3531    run.scr      zglasova     r     11/22/2004 11:25:06
3420    racun3       rvianell     r     11/14/2004 20:38:51
3532    submit2      bkovacev     qw    11/22/2004 11:32:57
```

Komponenta **Red poslova** prima zahtjeve za izvršavanjem poslova od korisnika, sprema poslove u redove i upravlja redovima poslova. Red poslova kontaktira komponentu Raspoređivač poslova i šalje podatke o poslovima koji čekaju u redovima. Korisnik pomoću Reda poslova dohvaća sve informacije o svojim poslovima. Dodatno, Red poslova sprema podatke o izvršenim poslovima (npr. količina memorije, trajanje posla) te različitih statistika o poslovima.

Komponenta **Raspoređivač poslova** odgovorna je za određivanje načina izvođenja poslova, tj. definiranje kada i gdje će se pojedini poslovi izvoditi. Pri tome Raspoređivač koristi tri skupa podataka: podatke o poslu (npr. zahtijevani broj procesora, količina memorije i HD), podatke o čvorovima klastera (npr. opterećenost čvorova, količina slobodne memorije i HD) i utvrđenih pravila, policy raspoređivanja. Podatke o poslovima dobiva od Reda poslova, a podatke o stanju i opterećenosti resursa od Upravitelja resursima. Politike raspoređivanje definira administrator klastera i na osnovu njih se definira politika korištenja klastera, tj. definira se koja vrsta poslova (npr. duži, kraći, usporedni, interaktivni) ima prednost. Raspoređivač poslova zadužen je da, u skladu s politikom raspoređivanja, optimira korištenje klastera.

Upravitelj resursima zadužen je za prikupljanje podataka o stanju čvorova te pokretanje i praćenje izvršavanja poslova. Upravitelj Resursima se sastoji od dvije komponente: poslužitelja i klijenata. Klijenti su servisi koji se izvršavaju na svim čvorovima klastera i zaduženi su za pripremanje okoline za izvršavanje poslova, praćenje stanja čvorova tepokretanje i praćenje izvođenjakorisničkih aplikacija.Poslužitelj je servis smješten na *frontendu* koji prikuplja informacije od klijenata, šalje poslove na izvođenje klijentima te pruža informacije komponentama Raspoređivač i Red poslova.

Osim SGE sustava za upravljanje poslovima, ovdje možete pronaći nešto više i o Torque i Maui sustavu.

Datotečni sustavi

Datotečni sustav je tip pohranjivanja i organiziranja podataka u spremnik podataka kojim se koristi operativni sustav. To je zapravo skup pravila I metoda pohrane podataka tako da operativni sustav u svakom trenutku jasno raspoznaje gdje se nalazi početak, a gdje kraj pohranjene informacije.

Na Srcu je u uporabi BeeGFS paralelni datotečni sustav. Takav sustav omogućava paralelno spremanje datoteka na više podatkovnih elemenata i učinkovit je za spremanje velikih datoteka.

Računalni klaster Isabella

Računalni klaster Isabella je nastao 2002. godine s ciljem da omogući svim zainteresiranim hrvatskim znanstvenicima pristup računalnom klasteru i rad na europskom projektu DataGrid kojeg vodi CERN.

Danas se Isabella sastoji od 135 računalnih čvorova koji ukupno sadržavaju 3100 procesorskih jezgri i 12 grafičkih procesora kao zajednički resurs svih znanstvenika u Hrvatskoj omogućava korištenje značajnih računalnih resursa pri zahtjevnim obradama podataka u sklopu znanstveno-istraživačkih projekata.

Dio čvorova povezan je u jedinstveni virtualni računalni sustav (Single System Image, SSI) pomoću sustava ScaleMP - ukupno 160 procesorskih jezgri i 2 TB radne memorije.

Tehničke karakteristike računalnog klastera Isabella možete pronaći na sljedećoj stranici: https://www.srce.unizg.hr/usluge/isabella/tehnicke-karakteristike.

Korisni linkovi

- Web stranice - novosti, tehničke informacije i dokumenti
    - http://isabella.srce.hr
- Web stranice za nadzor
    - http://teran.srce.hr/ganglia
- Služba pomoći - problemi s posredničkim sustavom, aplikacijama, pomoć u pripremi korisničkih aplikacija...
    - isabella@srce.hr