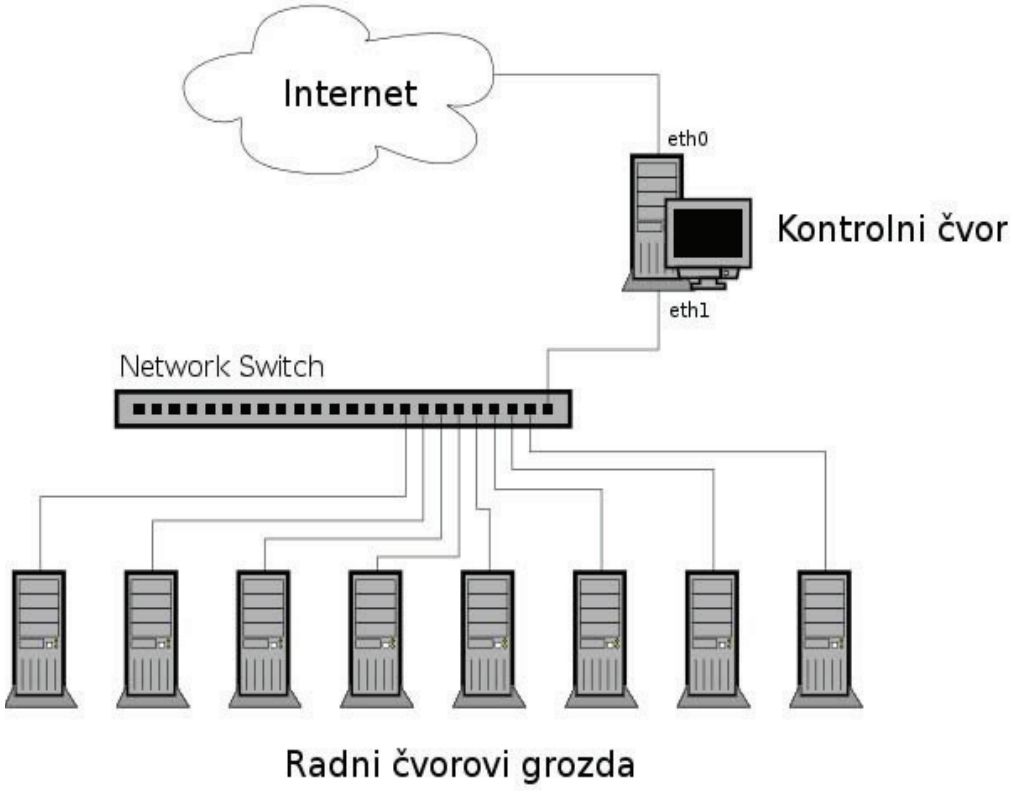
4 Osnove grozd (cluster) računala

4.1 Što je to računalni grozd (cluster)

Računalni grozdovi, ili clusteri, predstavljaju model uvezivanja više samostalnih PC računala za potrebe paralelnog računanja. Fizička topologija mreže u tipičnom računalnom grozdu prikazana je na slici.



Kontrolni čvor, kao što se vidi iz slike, jedini je čvor koji je direktno vezan na javnu mrežu, pa je jedna od njegovih osnovnih funkcija pružanje usluge udaljenog pristupa grozdu. Također, na kontrolnom čvoru pokrenuti su svi mrežni servisi potrebni za rad grozda kao integralne cjeline. Kako je glavni čvor na neki način najvažniji čvor u grozdu, važno je osigurati njegov neprekidan rad, jer će ispadom kontrolnog čvora iz funkcije, cijelo grozd računalo biti neupotrebljivo.

S druge strane, radna snaga računalnog grozda leži u radnim čvorovima. Radni čvorovi najčešće čine jaka, višeprocesorska računala, sa mnogo memorije, i njihova je jedina funkcija – izvršavanje (paralelnih) korisničkih aplikacija.

4.2 Što skupinu nezavisnih PC računala čini grozdom?

Da bi računala spojena kako je prikazano na gornjoj slici osposobili za paralelno računanje neophodno je, osim mrežnog povezivanja na njima osposobiti nekolicinu mrežnih servisa koji će omogućiti komforno administriranje i korištenje grozda kao integralne cjeline.

4.2.1 Mehanizam za automatsku instalaciju radnih čvorova

Broj radnih čvorova u računalnom grozdu može biti vrlo velik. Stoga je neophodno uspostaviti mehanizam koji omogućava automatsku instalaciju radnih čvorova, što će uvelike ubrzati proces postavljanja grozda u produkcijsko okruženje. Model automatske instalacije mora biti fleksibilan s obzirom na mogućnost konfiguracije, i mora omogućavati raznolikost konfiguracija za različite skupine radnih čvorova. Tako npr. neki radni čvorovi mogu imati instalirane aplikacije za kemiju, drugi za bioinformatiku, ili se pak čvorovi mogu razlikovati u konfiguraciji diskovnog sustava, za što će također biti potrebno omogućiti različite konfiguracije operacijskog sustava koji se instalira na konkretne radne čvorove.

4.2.2 Raspoređivač poslova (queueing sustav)

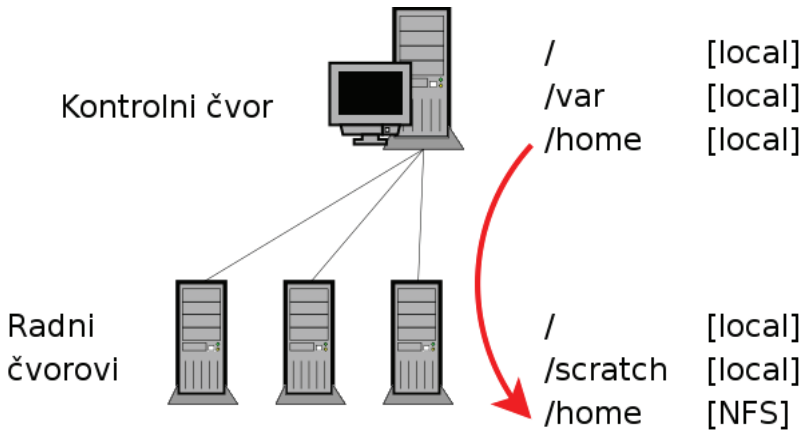
Queueing sustav je u grozd računalu zadužen za upravljanje resursima. Npr. ako neki korisnik želi pokrenuti svoju aplikaciju na 5 računala, taj će zahtjev predati queueing sustavu, te ako u datom trenutku u grozdu postoji 5 slobodnih računala, queueing sustav će taj posao pokrenuti. U protivnom, ako su svi čvorovi zauzeti, korisnički zahtjev bit će stavljen u red čekanja, te će kada zatraženi resurs bude raspoloživ, taj posao pokrenuti.

Queueing sustav se sastoji tri komponente, čija je funkcionalnost i mjesto izvođenja navedeno u slijedećoj tablici:

Naziv servisa	*Mjesto izvođenja*	*Opis funkcionalnosti*
Resource Manager Server	Kontrolni čvor	Glavni servis queueing sustava. Interakcija s korisnikom. Sadrži informacije o trenutnom zauzeću resursa unutar grozd računala. Inicira pokretanje korisničkog posla na radnim čvorovima. Inicira nasilno zaustavljanje korisničkih poslova ukoliko je došlo do prekoračenja zatraženih resursa.
Scheduler	Kontrolni čvor	Zadužen za informiranje Resource Managera o poslovima koji se nalaze u redu čekanja. Rukuje prioritetima, određuje koji će se posao slijedeći pokrenuti. Fair-share model raspolaganja resursima.
Job Executor	Radni čvor	Na zahtjev Resource Manager Servera, pokreće ili zaustavlja korisnički posao na radnom čvoru. Izvještava Resource Managera o trenutnoj zauzetosti resursa na radnom čvoru.

4.2.3 Distribucija korisničkih direktorija

Korisnici grozda pristupaju sustavu preko kontrolnog čvora, na kojem su fizički smješteni korisnički direktoriji. Međutim, korisnički poslovi koji će se izvršavati na radnim čvorovima najčešće ovise o ulaznim datotekama ili nekim bibliotekama koje kreira sam korisnik, pa je stoga nužno sadržaj korisničkih direktorija distribuirati na radne čvorove clustera. Na taj način, korisnički rad postaje mnogo komforniji, jer korisnik ne mora sam brinuti o preseljenju datoteka potrebnih za izvršavanje njegovog posla. Kako je pristup korisničkim direktorijima sa radnih čvorova najčešće ostvaren putem NFS (Network File System) mrežnog protokola, važno je imati na umu ograničenja koje postavlja takav model pristupa. Naime, ako korisnička aplikacija (koja se izvršava na radnom čvoru) obavlja masivno pisanje/čitanje na datotekama koje se nalaze u korisničkom direktoriju, obzirom da je pristup ostvaren preko mreže, može se dogoditi da izvršavanje aplikacije dovede do zagušenja mrežnog sustava unutar grozda, što će za posljedicu imati degradaciju performansi izvršavanja same aplikacije. U tom slučaju najbolje je posao koncipirati tako da se intenzivno pisanje/čitanje ulaznih/izlaznih datoteka obavlja na lokalnom disku u konkretnom radnom čvoru, te da se konačan rezultat konačno presnimi u korisnikov direktorij.

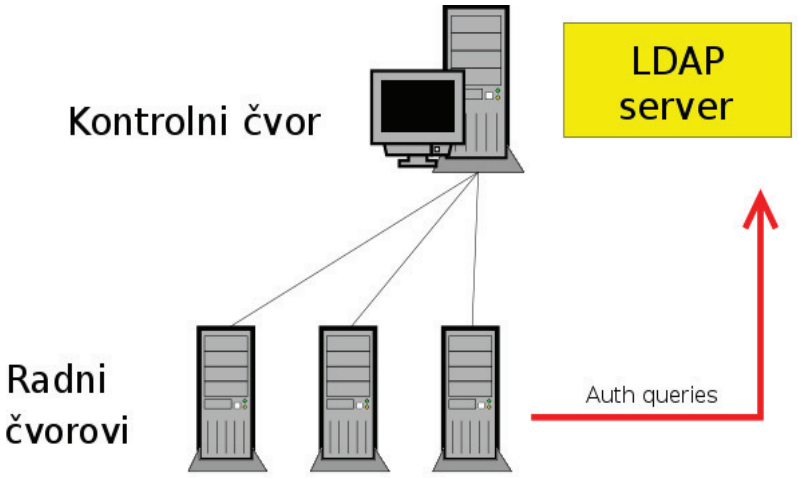


Možemo primjetiti da na radnim čvorovima imamo /scratch particiju. Ona nam prvenstveno služi kao medij za intenzivne čitaj/piši operacije unutar korisničkih aplikacija. Na taj način izbjegava se spomenuta degradacija performansi u slučaju intenzivnog čitanja/pisanja po udaljenom datotečnom sustavu.

4.2.4 Centralizacija upravljanja korisničkim računima

Prilično je jasno da ako računalni grozd sadrži veliki broj računala (može imati i preko nekoliko stotina), nužno je ugraditi neku vrstu centraliziranog upravljanja korisničkim računima. Za tu potrebu u nekim grozd-orjentiranim distribucijama koristi se NIS (Network Information Services), no u novije vrijeme LDAP (Lightweight Directory Access Protocol), zbog svoje fleksibilnosti i otvorenosti predstavlja sve interesantnije rješenje problema distribuiranog upravljanja korisničkim računima na računalima unutar grozdova.

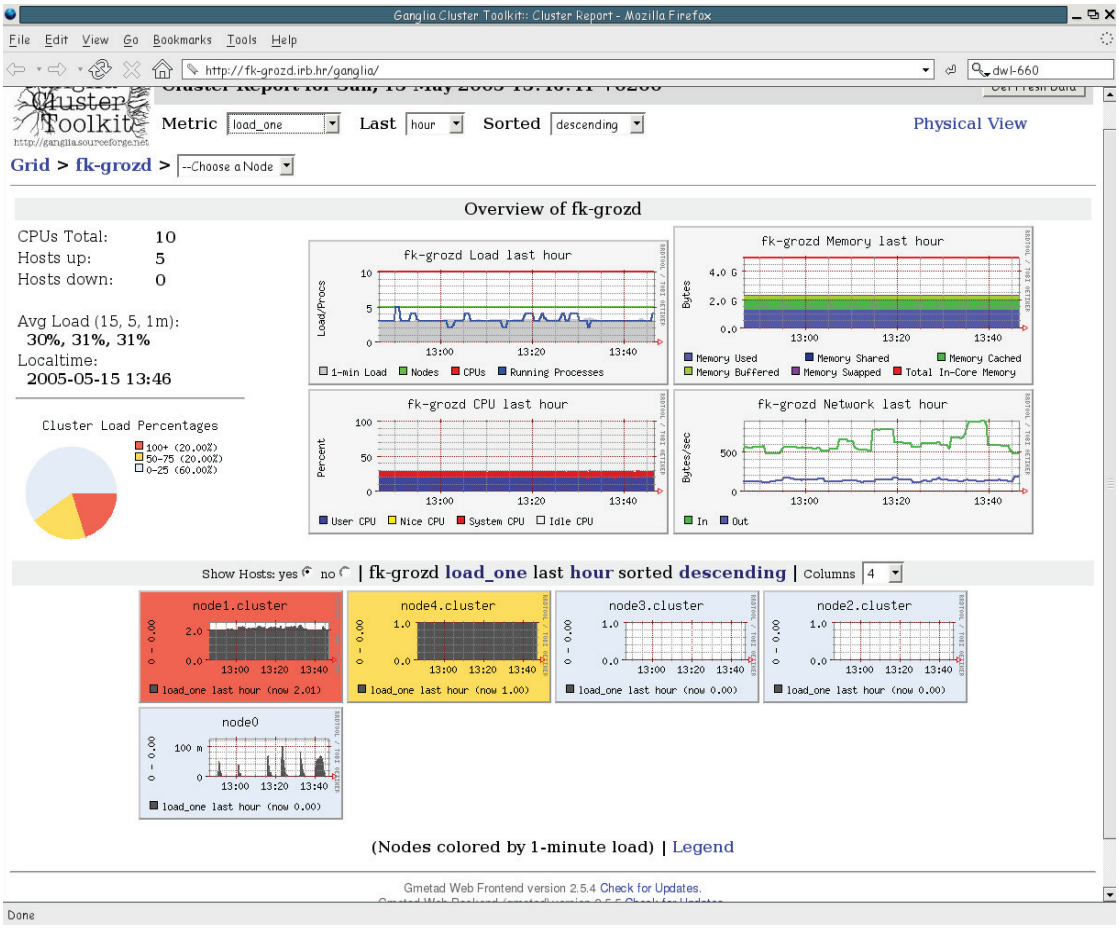
Kod ovakvog modela autentifikacije korisnika, baza podataka sa svim informacijama o korisniku (ime, grupa, sažetak lozinke, prioritet itd...) nalazi se na kontrolnom čvoru grozda, a kada se korisnik prijavljuje na radni čvor, tada autentifikacijski mehanizam na radnom čvoru formira upit prema centralnoj bazi na kontrolnom čvoru, te od tamo vuče sve informacije potrebne za provjeru autentičnosti korisnika.



Kako se ponekad upiti događaju vrlo intenzivno, sa velikog broja radnih čvorova i od velikog broja korisnika, poželjno je na radnim čvorovima imati “cacheing” sustav koji će privremeno spremati informacije iz LDAP poslužitelja i kasnije ih koristiti umjesto ponovnog postavljanja upita prema kontrolnom čvoru. Na taj se način izbjegava opterećivanje mrežnog sustava, što za posljedicu ima ukupno poboljšanje performansi cijelog sustava.

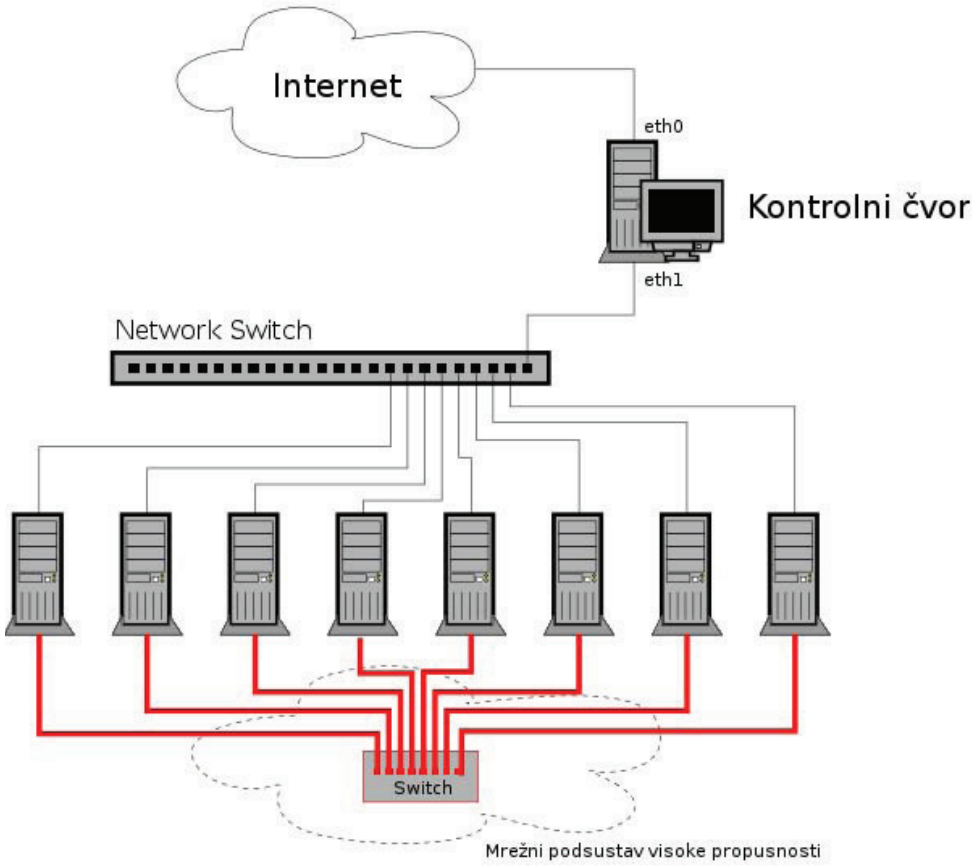
4.2.5 Centralizirani nadzor opterećenja sustava

Zbog preglednosti trenutne situacije na grozdu računala, praktično je imati pregledan grafički prikaz opterećenja sustava za sve čvorove, a poželjno je da je takav sustav proširiv kako bi se mogle mjeriti različite veličine, a ne samo ponuđene. Najčešće se u računalnim grozdovima koristi sustav nadzora pod imenom “Ganglia”. Primjer prikaza opterećenosti računalnog grozda pomoću Ganglia sustava možemo vidjeti na slici dolje.



4.2.6 Specijalizirana mrežna sučelja visoke propusnosti

Kod paralelnih aplikacija koje imaju vrlo intenzivnu međuprocesnu komunikaciju, usko grlo čini upravo mrežno sučelje preko kojeg se ta komunikacija odvija. Povećanje mrežne propusnosti, a time i performasi izvođenja paralelne aplikacije može se ostvariti dodavanjem specijalnog mrežnog podsustava namijenjenog isključivo međuprocesnoj komunikaciji.

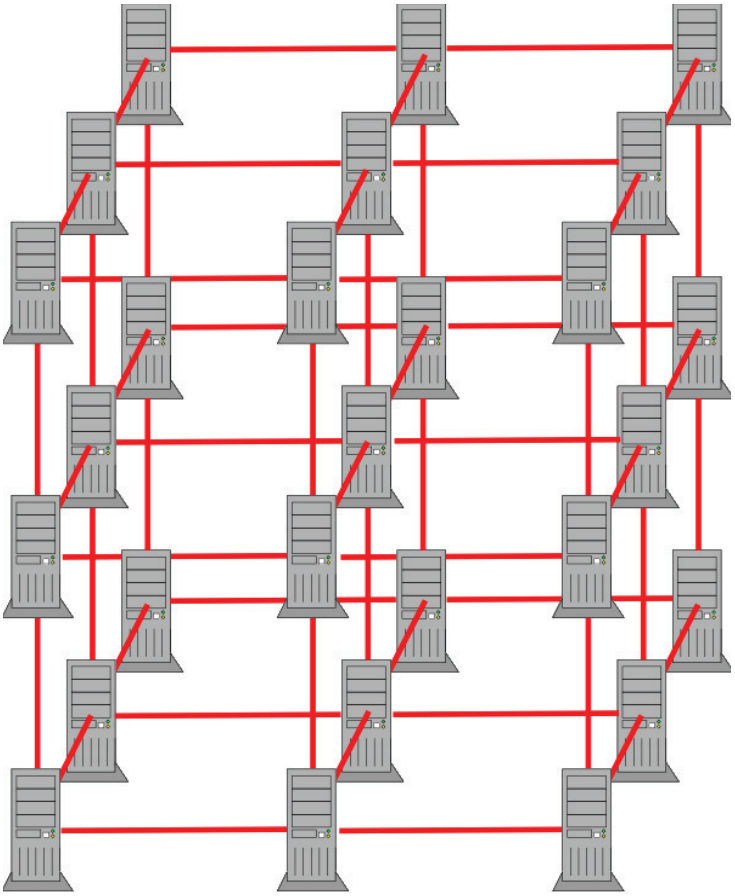


Svakom radnom čvoru ugrađuje se dodatno (najčešće optičko) mrežno sučelje prema mrežnom preklopniku (switch), čija je brzina višestruko veća od standardnih mrežnih podsustava.

Važno je napomenuti da specijalizirana mrežna sučelja omogućavaju optimizaciju komunikacije u paralelnim aplikacijama na još jedan način. Naime, kako standardno mrežno sučelje funkcionira na IP protokolu, na svaku poruku koja se šalje između dva procesa potrebno je dodati zaglavlje IP protokola, koje je ponekad puno veće od samog sadržaja poslane poruke. Kako bi se izbjeglo nepotrebno opterećivanje mreže dugačkim zaglavljima IP protokola, cijela međuprocesna komunikacija se na specijaliziranom mrežnom podsustavu može ostvariti na nižem nivou, posve izostavljajući pritom IP protokol. Takva međuprocesna komunikacija ostvariva je jedino posebno pripremljenim bibliotekama za paralelno programiranje u kojima je uključena podrška za ovakav način rada.

4.2.7 Specijalizirana mrežna topologija

Uvođenje specijaliziranog mrežnog podsustava u računalni grozd svakako donosi poboljšanje performansi izvođenja aplikacije, međutim ponekad je zbog dodatnih optimizacija komunikacije potrebno mrežnu topologiju grozda prilagoditi konkretnoj aplikaciji. Npr. kod problema simulacije atmosferskih prilika, podjela ukupnog problema ostvarena je podjelom simuliranog prostora na pravokutnike, između kojih se u svih šest smjerova događa komunikacija približno sličnog intenziteta. Za takvu vrstu problema, optimalna mrežna topologija spajanja radnih čvorova u takvu vrstu 3D rešetke prikazana je na slici.



4.3 Prednosti računalnih grozdova pred konvencionalnim superračunalima

Računalni grozd, kao model uvezivanja pojedinačnih PC računala relativno male snage u jedno virtualno superračunalo imaju nekolicinu prednosti pred velikim, monolitnim, konveniconalnim superračunalima. Iako je za veliku većinu paralelnih aplikacija pogodna primjena i jednog i drugog modela, računalni grozd svojom odnosom cijene i performansi znatno je ispred konvencionalnih superračunala. Osim cijene, usporedba nekih karakteristika računalnog grozda sa konvencionalnim superračunalom prikazana je u tablici.

	Računalni grozd	***Monolitno superračunalo***
Operacijski sustav	Linux, jednostavna instalacija.	Specijalizirani, komercijalni OS, izuzetno kompleksne instalacijske procedure.
Proširivost	Dodavanje novih PC računala niske cijene, dobavljivih u trgovinama. Mogućnost izbora komponenti različitih proizvođača.	Ovisnost o proizvođaču, visoka cijena specijaliziranih modula za konkretan tip superračunala
Ekologija	Računala ranije korištena u računalnom grozdu, kad postanu zastarjela za računanje, jednostavno se mogu prenamijeniti, i koristiti još određeno vrijeme.	Zastarjelo superračunalo obično je moguće jedino zamijeniti novim, modernijim i vjerojatno skupljim modelom.
Održavanje	Sve komponente sustava dobavljive su u trgovinama, cijena je prihvatljiva.	Ovisnost o proizvođaču, izuzetno visoka cijena rezervnih dijelova.
Programska podrška	Veliki broj open source i komercijalnih paralelnih aplikacija, razvojnih alata i kompilatora za paralelno okruženje.	Obzirom na egzotičnost platforme, ograničena fleksibilnost na primjenu različitih razvojnih alata, paralelnih aplikacija i kompilatora.