



UNIVERZITET SINGIDUNUM

**DEPARTMAN ZA POSLEDIPLOMSKE STUDIJE I
MEĐUNARODNU SARADNJU**

**- MASTER STUDIJSKI PROGRAM -
SAVREMENE INFORMACIONE TEHNOLOGIJE**

**IZGRADNJA SISTEMA VISOKIH PERFORMANSI
ZA PARALELNO RENDEROVANJE VIDEO
MATERIJALA UPOTREBOM RAČUNARSKE
MREŽE I KLASTERA**

– Master rad –

Mentor:

Prof. dr Mladen Veinović

Student:

Pavle Dakić

2014 410 081

Izgradnja sistema visokih performansi za paralelno renderovanje video materijala upotrebom računarske mreže i klastera

Sažetak: U ovom radu se proučava način izgradnje sistema visokih performansi. Razmatraju se finansijski, hardverski i softverski zahtevi potrebni za realizaciju istog. Fokus je na pristupima različite realizacije sistema visokih performansi, koji za cilj imaju integraciju i povezivanje većeg broja računara, upotrebom različitih mrežnih inter konekcija. Vršiti se analiza postojećih hardverskih i softverskih rešenja iz ove oblasti, kao i njihova isplativost. Na osnovu tabelarnog prikaza biće izvršena analiza, brzine izračunavanja CPU i GPU renderovanja i njihovog dolaska do finalnog proizvoda, tj. rešenja. Razlog za upotrebu klaster HPC – sistema visokih performansi je u mogućnostima izvršavanja kalkulacija za više različitih naučnih oblasti.

Ključne reči: Sistemi visokih performansi, grid računarstvo, računarstvo u oblaku, računarske mreže, klaster, renderovanje video materijala.

Building a high-performance systems for parallel rendering of video material using computer networks and clusters

Abstract: In this master paper is described the way of building high-performance systems. What are the financial, hardware and software requirements needed for its execution. The focus is on different approaches to the implementation of high-performance, aimed at integrating and connecting multiple computers, using different network inter connection. Performing analysis of existing hardware and software solutions in this area, as well as their profitability. Based on the tabulation data, data will be analyzed, showing calculating speed of CPU and GPU rendering and their needed time for arrival to the final stage, getting the solution. The reason for use of HPC cluster - the high-performance capabilities is in performing various calculations for different scientific fields.

Keywords: High performance systems, grid computing, cloud computing, computer networks, clusters, rendering of video material.

SADRŽAJ:

Uvod	5
1 Metodologija istraživačkog rada	7
1.1 Uvodne napomene i obrazloženje rada	7
1.2 Predmet istraživanja	7
1.3 Ciljevi i zadaci istraživanja	7
1.4 Istraživačke hipoteze	9
1.5 Metodi istraživanja i tok istraživačkog procesa	9
2 Kompanije koje proizvode HPC sisteme	10
2.1 HP, Hewlett-Packard	13
2.1.1 HP Moonshot	13
2.1.1.1 Moonshot 1500	13
2.1.1.2 HP ProLiant m710	16
2.1.2 HP Apollo 8000 sistem	17
2.1.2.1 HP Apollo f8000 rack	18
2.1.2.2 HP Apollo 8000 iCDU Rack	19
2.1.2.3 HP ProLiant XL730f server	23
2.1.2.4 HP Infiniband switch	24
2.1.3 Usporedni prikaz karakteristika HP	27
2.2 IBM	28
2.2.1 IBM Blue Gene	28
2.2.1.1 Analiza JuQueen konfiguracija	33
2.2.2 IBM Power 775	34
2.2.3 Usporedni prikaz karakteristika IBM	36
2.3 Cray Inc.	37
2.3.1 XC SERIES	37
2.3.1.1 Cray XC40	38
2.3.1.2 Cray XC40-AC	41
2.3.2 Usporedni prikaz karakteristika Cray XC	42
3 Protokoli za komunikaciju HPC sistema	44
3.1 MPI	46
3.1.1 OpenMPI	46
3.2 Interconnect	47
3.3 InfiniBand	48
4 Računarska mreža	52
4.1 HPC topologija	52
4.1.1 Fat tree topologija	52
4.2 Klaster	55
4.3 Balansiranje saobraćaja	57
4.3.1 DNS balansiranje saobraćaja	57
5 Operativni sistemi HPC	59
6 Laboratorijsko okruženje	60
6.1 Izgradnja HPC klastera - Build	61
6.2 Izbor odgovarajućih servisa - Rolls	61
6.3 Informacije o klasteru	62
6.4 Podešavanje mreže	63
6.5 Gateway, DNS server	64
6.6 Root Password	64
6.7 Vremenska zona, NTP server	65
6.8 Automatsko Partitionisanje i formatiranje hard diska	65
6.9 Ručno Partitionisanje i formatiranje hard diska	66
6.10 Instalacija klastera	67
6.11 Logovanje na sistem	67
6.12 Logovanje na sistem, upozorenje	68
6.13 Dodavanje novih računskih nodova	68
6.14 Dodavanje novih računskih nodova, opcija Compute	69
6.15 Dodavanje novih računskih nodova, ostvarivanje komunikacije	69
6.16 Ganglia	70

6.17 Virtualizacija sistema.....	72
Zaključak.....	73
Literatura	75
Spisak slika.....	78
Spisak tabela	80
Spisak formula	80

Uvod

Živimo u digitalnom svetu gde smo svi povezani na Internet, na određeni način posredstvom jednog ili drugog uređaja, zbog čega uspešnost pozicioniranja jedne kompanije na tržištu, mnogo zavisi od brzine plasiranja informacija različitog oblika (teksta, slike, videa itd.) od momenta kada se neki događaj završio, jer većina današnjih digitalnih korisnika ima zahteve po principu “sad i odmah”, to jest zahteva se praćenje događaja u realnom vremenu upotrebom različitih uređaja. Jer savremena publika je sve zahtevnija po pitanju kvaliteta digitalnog sadržaja. Zbog čega svi podaci, informacije i sadržaji moraju da se digitalizuju, pre nego što se prikažu krajnjem korisniku.

Računarstvo visokih performansi obuhvata proces naprednog matematičkog računanja korišćenjem paralelnog procesiranja, omogućavajući brže izvršavanje intenzivnih računarskih operacija kao što su krypto analiza, istraživanja klimatskih promena, rudarenje podataka (eng. data mining), obradu velike količine podataka (eng. big data), dizajnirane i modelovanje automobila i aviona, molekularno DNK modelovanje, berze, naftna industrija itd. [1] Svaki digitalni podatak u računaru je predstavljen kao kolekcija bitova, koja je sačuvana u određenom file formatu, koji sadrži određenu strukturu, način kompresovanja, čitanja, pisanja itd. File formati su neophodni kako bi bila moguća obrada podataka u cilju izvršavanja operacija poput: arhiviranja, emitovanja sadržaja preko računarskih mreža i na kraju prikazivanja na ekranima. Multimedijalni sadržaj je pre svega neophodno proizvesti, a onda preneti do krajnjeg korisnika, upotrebom jednog od dva načina: online (upotrebom računarske mreže) ili offline (upotrebom memoriskih uređaja, SD, CD, DVD, SSD itd.). Važnost video materijala kao jedne digitalne celine, kojom se prikazuje struktura informacija je u postojanosti svih njenih elementa od muzike, teksta, slike itd. Zahvaljujući integraciji sadržaja moguće je istovremeno aktiviranje više čula, sa percepcijom vizuelnih i slušnih sposobnosti, ostvarujući realniji doživljaj prikazane digitalne celine.

Video obrada uključuje snimanje, transkodiranje i distribuciju (eng. streaming) video materijala, tako da sam proces snimanja podrazumeva pretvaranje analognog video signala u digitalni, upotrebom video kamere ili bilo kog drugog uređaja, koji vrši generisanje signala. Dobijeni digitalni podaci se predstavljaju u vidu fajlova koji su kodirani u određenom video formatu, koji se kasnije koriste za dalju obradu. U slučaju prenosa uživo (eng. live streaming), gde se fajlovi najčešće prenose upotrebom mrežnih tehnologija, vrši se direktno video transkodiranje i kompresovanje izvora signala, upotrebom HPC sistema, pre nego što se isporuče krajnjem korisniku.

Obrada velike količine informacija i dobijanje informacija na što brži i efikasniji način, tačnije u realnom vremenu zahteva postojanost određene računarske infrastrukture, u vidu prostora za fizički smeštaj servera, računarske mreže, hardvera, softvera i stručnog osoblja koje će vršiti administraciju i podešavanje servera. Renderovanje kao jedan složen proces obrade digitalnih podataka, zahteva intenzivnu upotrebu velikog broja računara i njihovih resursa, za potrebe izračunavanja i dobijanja finalnog rezultata u što kraćem vremenskom periodu, tako da vremenski proces renderovanja po individualnom frejmu (eng. frame), može da varira od nekoliko sekundi, pa sve do nekoliko sati, dana, meseci ili godina; u zavisnosti od računarske snage na raspolaganju, to jest broja CPU i GPU koji rade zajedno unutar klastera (eng. cluster). [2] [3]

Vreme renderovanja najviše zavisi od kompleksnosti scene, korišćenih specijalnih efekata, dubine i upotrebe senki, svetlosti i drugih efekata. Na primer, kratka animacija može da bude trajanja 2-3 minuta, i da koristi 30 frajmova u sekundi (eng. Frames per second – fps) i da za nju bude potrebno **nekoliko dana** za dobijanje finalnog rezultata. Dok za renderovanje scene sa tekstom, bez specijalnih efekata može biti potrebno samo **nekoliko minuta**. Današnji super računar koji se koristi za realizaciju i izvršavanje svih ovih složenih

operacija se ne razlikuje mnogo od standardnog računara. Karakteristike koje ga izdvajaju su brzina procesora i ostalih komponenti, koje su daleko iznad svih trenutno postojećih računara. Svet računarstva i elektronike se konstantno menja i razvija, zbog čega većina najbržih super računara i HPC sistema nosi svoje laskave titule samo nekoliko godina u najboljem slučaju. Po rečima inženjerima i programera koji su aktivno uključeni u razvoj i realizaciju novih tehnologija, računari koji su danas super računari, sutra će biti u većini kompanija standardna radna stanica (eng. workstation). [4] [5]

Konstantan rast broja korisnika i količine saobraćaja koji se generiše, doveo je do ubrzanog razvoja usluga računarstva u oblaku, mrežnih servisa, data centara, komunikacionih protokola, video sadržaja, servera, tehnologija itd. postavljajući nove komunikacione standarde i brzinu mrežne komunikacije sa minimalnim propusnim kapacitetom od 1 Gbps. Klasteri se sastoje od jednog glavnog čvorišta i niza računskih nodova koji su fokusirani na izvršavanje računskih zadataka. U najjednostavnijoj konfiguraciji, pristup resursima je najvažnija usluga o strane glavnog čvora za ceo klaster koji mora da poseduje mogućnost opsluživanja velikog broja korisnika.

Istraživanja u nauci i rešavanja različitih opterećenja, kao što su modelovanje predmeta, klimatskih promena, simulacija vetra, animacija itd. zahteva postojanost određenih fizičkih i hardverskih uslova. Na primer jedna složena elektromagnetna video simulacija zahteva 7.8 teraflopsa i 12 sati rada za renderovanje, to jest dobijanje jednog "komada" simulacije. Nije slučajno što se HPC sistemi procenjuju na nekoliko miliona dolara ili evra, u zavisnosti od broja ostvarenih teraflopsa. Izvršavanje simulacija zahteva najefikasnije računarske platforme. Vreme potrebno za izvršavanje jedne simulacije zavisi od mnogo faktora, to jest broja CPU/GPU jezgara na raspolaganju i njihov faktor iskorišćenja u cilju dobijanja efikasnosti i skalabilnosti usled obavljanja Interconnect operacija unutar klastera. Efikasni računarski sistemi visokih performansi zahtevaju postojanost širokog propusnog opsega, to jest minimalnu 1Gbps mrežnu konekciju, malo kašnjenje (eng. latency) između hiljade procesorskih čvorova i veliku brzinu sistema za skladištenje podataka na mreži ili lokalnom računaru. Za skladištenje podataka se većinom koriste SAN i NAS sistemi, zbog svojih mogućnosti upotrebe specijalizovanih fajl sistema. [6] [7]

Kako bi se jasnije shvatio ceo koncept, potrebno je sagledati trenutno stanje raspoloživih tehnologija i rešenja. Na tržištu svaka kompanija želi da bude konkurentna zbog čega je svaka razlika bitna, od kvaliteta proizvoda/usluge do krajnje isporuke, da baš njihov proizvod bude izabran od strane krajnjeg korisnika, zbog čega ćemo uraditi analizu kompanija i njihovih ponuda, jer neke kompanije će imati znatno lošiji marketing, a znatno kvalitetniji i povoljniji proizvod za razliku od drugih kompanija koje će gledati samo krajnji profit i ispunjenje minimalnih tržišnih zahteva, u cilju razumevanja komponenti koje se koriste prilikom izgradnje i sastavljanja celokupnog sistema visokih performansi stavljajući akcenat na komunikacione protokole i njihove mogućnosti primene.

1 Metodologija istraživačkog rada

1.1 Uvodne napomene i obrazloženje rada

Cilj ovog rada je opisivanje i analiza potrebnih resursa za izgradnju jednog sistema visokih performansi za različite naučne potrebe na fakultetima, univerzitetima i drugim ustanovama, koje imaju potrebe za aktivnom upotrebom mogućnosti HPC sistema kao i svima, koji žele da na što brži i efikasniji način dođu do finalnog proizvoda/rešenja pružajući osnovu za izgradnju sistema od strane manjih kompanijama i samih pojedinca, koji se tek upoznaju sa ovom oblašću pružajući razumevanje komunikacionih protokola i njihovih mogućnosti sa različitim implementacijama unutar operativnih sistema. Postavljajući osnovnu strukturu HPC sistema prilikom izbora hardverskog rešenja pružajući korisnicima bolji pristup i sagledavanje trenutnih HPC operativnih sistema.

1.2 Predmet istraživanja

Predmet istraživanja je izgradnja sistema visokih performansi za paralelno renderovanje video materijala upotrebom računarske mreže i klastera, usmerena na proces pronalaska primenljivih rešenja za obradu digitalnog video materijala, koja poseduju mogućnost primene određenih protokola i mogućnosti HPC sistema analizujući postojeće protokole i njihove implementacije unutar operativnih sistema. U cilju ostvarivanja maksimalne komunikacione brzine izražene u MBps, GBps, TBps itd

Ključni segmenti u okviru ovog rada se odnose na sledeće stavke:

1. Analiza tržišta i upotreba HPC sistema
2. Hardver, fizički serveri
3. Naučna namena i primena postojećih HPC sistema
4. Operativni sistemi koji se koriste za realizaciju HPC sistema
5. Pregled hardverskih rešenja i softvera korišćenog za realizaciju

1.3 Ciljevi i zadaci istraživanja

Naučni cilj istraživanja je razumevanje i analiza postojećih rešenja sa njihovim mogućnostima koji podrazumeva upotrebu jednog ili više računara sa istim ili različitim hardverskim komponentama, u cilju postizanja određene brzine izračunavanja izražene u tera-flopsima (eng. Teraflops).

Konkretan cilj istraživanja je pronalazak odgovora na sledeća pitanja:

1. Cena i isplativost izgradnje sistema (prostor, struja, hlađenje itd).
2. Koliko kompanija se bavi proizvodnjom HPC sistema.
3. Potrebno vreme za renderovanje jednog složenog (3D scena) materijala upotrebom ovog sistema.
4. Najkorišćeniji protokoli koji se koriste prilikom komunikacije HPC sistema.
5. Mogućnost propusnosti podataka i balansiranja saobraćaja unutar računarske mreže.
6. Implementacija protokola, brzine i upotreba.
7. Realizacije topologije HPC sistema.
8. Operativni sistemi koji se koriste za realizaciju HPC sistema i koji sadrže različite implementacije komunikacionih protokola.
9. Softver koji može da koristi klaster (eng. Cluster), mrežnu infrastrukturu HPC sistema za potrebe renderovanja različitog video materijala.

Glavni cilj ovog master rada je razumevanje načina funkcionisanja HPC sistema, u cilju izrade realne simulacije unutar virtuelnog okruženja, pre izgradnje sistema na fizičkim serverima razmatrajući operativne sisteme i njihove mogućnosti realizacije klastera, prilikom konfiguracije, instalacije, podešavanja itd. Simulacija će podrazumevati osnove sistema visokih performansi, to jest kreiranje većeg broja računara unutar virtuelnog okruženja, sa različitim ili istim operativnim sistemima i njihovim softverskim rešenjima, koji će zajedno činiti klaster (eng. cluster).

1.4 Istraživačke hipoteze

Opšta hipoteza

Informacije i podaci koji se obrađuju i prenose posredstvom različitih komunikacionih medija, a zatim nakon obrade smeštaju na jedan ili više različitih servera unutar mreže, moraju da poseduju sigurnu i neprekidnu mrežnu komunikaciju, kako podaci ne bi mogli da budu kompromitovani ili ukradeni, zbog čega je potrebno obraditi pažnju na sigurnost i zaštitu podataka, koji se razmenjuju unutar ovog sistema. Da bi se izvršila konfiguracija neometane i sigurne komunikacije, neophodno je dobro poznavanje oblasti: zaštite informacija i oblasti računarskih mreža, protokola koji se koriste za komunikaciju između servera i klijenata, digitalne forenzike, geografskih informacionih sistema itd.

Radna hipoteza

Nakon pregleda specifikacija rešenja biće izvršena analiza par najkvalitetnijih, koji mogu da se koriste u laboratorijskom okruženju, gde će biti izvršeno testiranje i analiza dobijenih rezultata. Svrha analize je u dobijanje rezultata na osnovu brzine izračunavanja, kako bi se uvidele prednosti i nedostaci određenog rešenja. Biće izvršena analiza implementacije protokola unutar operativnih sistema. Svrha analize je razumevanje propusnosti određenih komunikacionih protokola, u cilju lakšeg izbora odgovarajućeg softverskog i hardverskog rešenja

1.5 Metodi istraživanja i tok istraživačkog procesa

Konkretna metodologija istraživanja svodi se na upotrebu materijala prikupljenog sa Internet izvora, elektronskih biblioteka, video materijala, štampane literature iz oblasti: računarskih mreža, zaštite informacija, virtualizacije, operativnih sistema, obrade digitalnog signala (slike i video), protokola za komunikaciju kao i drugih povezanih oblasti. Metodom analize unutar laboratorijskog okruženja će se vršiti prikupljanje podataka i njihovih najvažnijih karakteristika. Upotrebom tabelarnog prikaza i metode komparacije prikupljenih podataka, dolazi se do izbora optimalnog sistema i mogućnosti za realizaciju eksperimenta, unutar virtuelnog laboratorijskog okruženja.

2 Kompanije koje proizvode HPC sisteme

Globalna ekonomija sa porastom prodaje računara i HPC sistemima, ima dobro predviđanje za budućnost. Razlog za to je povećana prodaja, ekoloških sistema (eng. green systems) u obliku servera, uređaja za skladištenje podataka, mrežne opreme i itd. što samim tim podrazumeva, jače procesore i veću potrošnju električne energije za izvršavanje složenijih operacija. Svaki današnji procesor je brži od super računara pre 10 godina, ali i dalje nedovoljno brz za ispunjavanje trenutnih zahteva.

Konstantni apetit za računarskom snagom u video industriji i boljim prikazima grafičkih sadržaja, zahteva upotrebu više procesora. Upotreba više procesora za dobijanje računarske snage je navela proizvođače, da prave manje radne stanice koje kasnije povezuju u računarsku mrežu (1-100Gbps), formirajući “mrežu radnih stanica“ (eng. network of workstations - NOW). Mreža radnih stanica (NOW) može se koristiti individualno, upotrebom serija (eng. batch) komandi ili upotrebom više računara, upotrebom protokola za komunikaciju i prosleđivanje poruka kao što su paralelna virtualna mašina – PVM ili interfejs za prosleđivanje poruka – MPI. [8]

Porast efikasnosti i prepoznatljivosti neke kompanije jeste posedovanje HPC sistema, zbog čega proizvođači HPC sistema moraju da ispune neke od *izazova* i *zahteva* korisnika.

Glavni izazovi za kompanije:

1. Napajanje, hlađenje, nekretnine (prodajni prostor), menadžment sistema.
2. Softverske prepreke, troškovi programiranja za različite arhitekture.
3. Pronalazak odgovarajuće procesorske snage.
4. Opsluživanje povećanog broja korisnika.
5. Menadžment uređaja za čuvanje podataka, i samih podataka koji su u porastu.

Ključni zahtevi korisnika:

1. Cena i performanse.
2. Ponuda proizvoda, kvalitet usluge i ukupni troškovi kupovine.
3. Integracija aplikacija, šta mogu da ponude kao osnov, iskustvo upotrebe HPC.
4. Dobra reputacija, tihi serveri, dobar ugovor o održavanju, istorija o servisiranju.
5. Tehničke inovacije, mogućnost uključivanja i razumevanja potreba u više godišnjem partnerstvu.
6. Razlika u ceni, ukupne podrške i servisiranja.

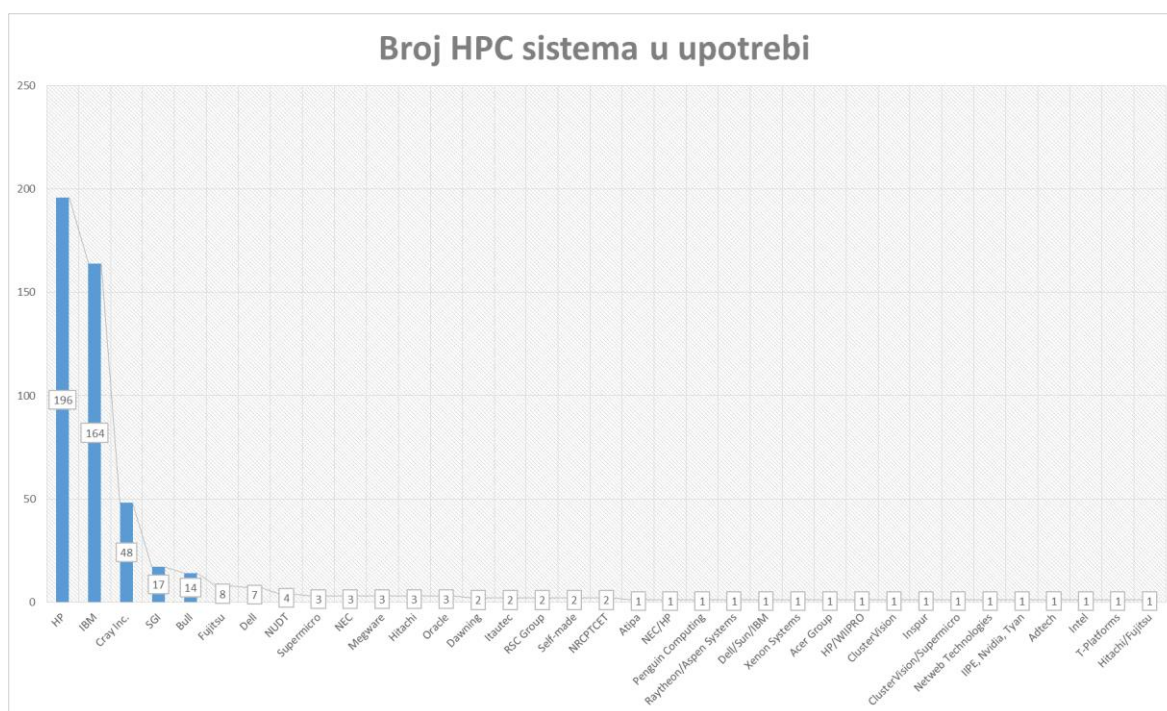
Selekciju i dalje istraživanje proizvođača izvršićemo na osnovu pregleda tržišta po sledećim kriterijumima:

1. Broj sistema sa kojima raspolažu.
2. Sistemski udeo.
3. Broj sistema u upotrebi.

Tabela 1. Proizvođači i njihovi sistemi, sa sistemskim udelom.

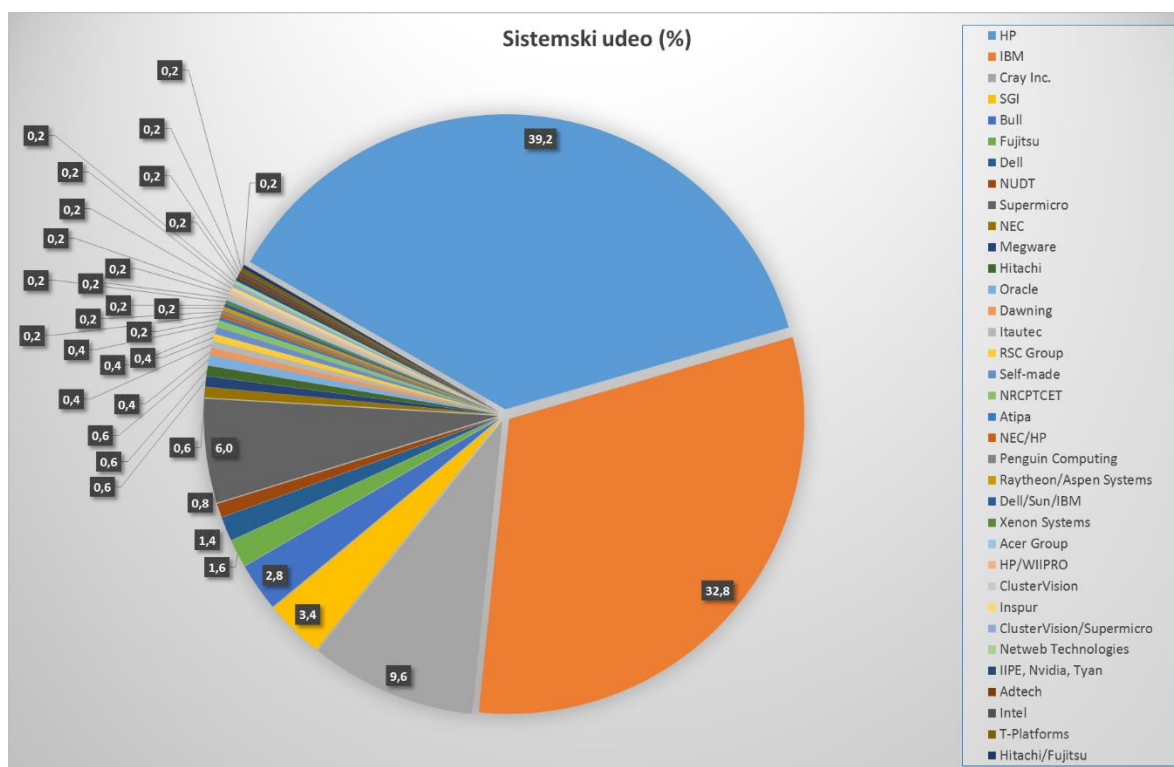
Proizvođač	Broj	Sistemski Udeo (%)
HP	196	39,2
IBM	164	32,8
Cray Inc.	48	9,6
SGI	17	3,4
Bull	14	2,8
Fujitsu	8	1,6
Dell	7	1,4
NUDT	4	0,8
Supermicro	3	6,0
NEC	3	0,6
Megware	3	0,6
Hitachi	3	0,6
Oracle	3	0,6
Dawning	2	0,4
Itautec	2	0,4
RSC Group	2	0,4
Self-made	2	0,4
NRCPTCET	2	0,4
Atipa	1	0,2
NEC/HP	1	0,2
Penguin Computing	1	0,2
Raytheon/Aspen Systems	1	0,2
Dell/Sun/IBM	1	0,2
Xenon Systems	1	0,2
Acer Group	1	0,2
HP/WIIPRO	1	0,2
ClusterVision	1	0,2
Inspur	1	0,2
ClusterVision/Supermicro	1	0,2
Netweb Technologies	1	0,2
IIPE, Nvidia, Tyan	1	0,2
Adtech	1	0,2
Intel	1	0,2
T-Platforms	1	0,2
Hitachi/Fujitsu	1	0,2

Izvor: sopstvena izrada, podaci preuzeti sa Top500.org, Decembar 2014.



Slika 1. Broj HPC sistema u upotrebi

Izvor: sopstvena izrada, podaci preuzeti sa Top500.org, Decembar 2014.



Slika 2. Sistemska udeo upotrebe HPC sistema

Izvor: sopstvena izrada, podaci preuzeti sa Top500.org, Decembar 2014

Na osnovu izvršene analize tržišta obradićemo 3 proizvođača koji imaju najveći broj sistema u upotrebi: **HP, IBM, Cray Inc.** i za svaku kompaniju ćemo izvršiti već navedenu analizu: cene, brzine izračunavanja, potrošnje električne energije, hardverske karakteristike, potreban fizički prostor za smeštaj server itd.

2.1 HP, Hewlett-Packard

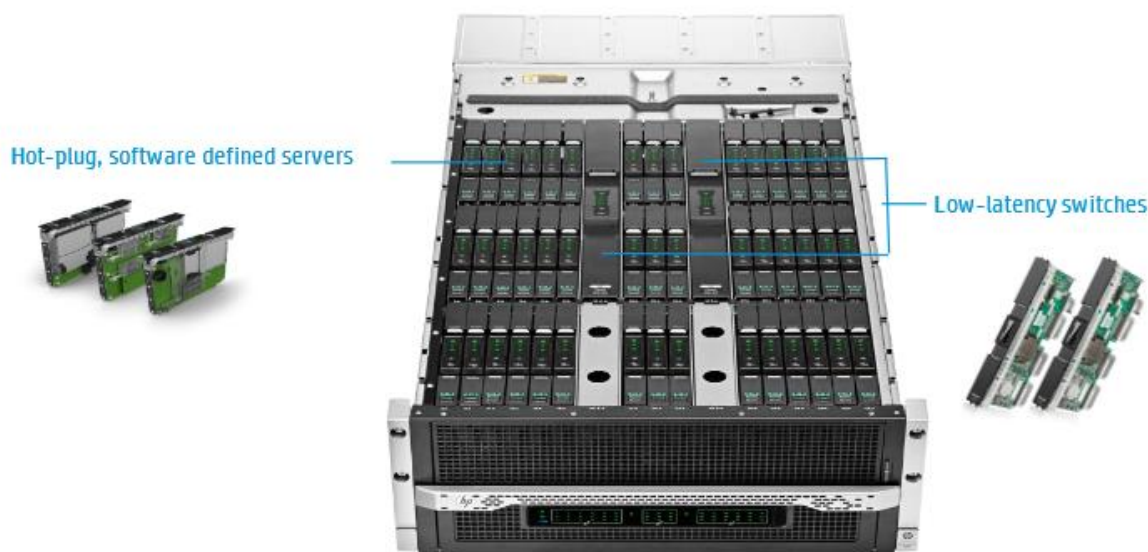
HP je multinacionalna kompanija za informacione tehnologije sa sedištem u Palto Alto, Kalifornija, Sjedinjene Američke države. Predstavlja jednu od vodećih kompanija na tržištu, u oblasti računarstva i tehnologija. HP kao lider u HPC rešenjima konstantno ulaže u istraživanja i stvaranje novih tehnologija. Trenutno u svojoj ponudi HPC sistema poseduje dva sistema Moonshot i Apollo, koji sadrže više različitih servera u zavisnosti od namene HPC sistema.

2.1.1 HP Moonshot

Moonshot predstavlja sistem modularnog pristupa, ugradnje komponenti. Koristi novu inovativnu arhitekturu jednostavnog novel dizajna, koji uključuje gust smeštaj servera pozicionirajući serverske module u serijski niz, u vidu kertridž slotova pružajući maksimalne performanse HPC sistema. Unutar jednog kućišta Moonshot servera, može se smestiti 45 HP ProLiant m710 servera, čija je širina 4.3U i visina 7.5 inča po kućištu, što samim omogućava veći broj servera po jednom kvadratnom metru povećavajući iskorišćenost postojeće infrastrukture, sa upotrebom manjeg prostora i električne energije za potrebe hlađenja i napajanja servera. [9] [10] [11]

2.1.1.1 Moonshot 1500

Serversko kućište Moonshot 1500 pruža modularnu snagu i mogućnost proširenja, u zavisnosti od potreba, minimizira kompleksnost instalacije, upravljanja i ispravljanja grešaka. Sistem pruža mogućnost hostovanja 45 nezavisnih ProLiant servera, koji su povezani na različite interfejsse. [12]



Slika 3. Moonshot 1500 kućište

Izvor: [13]

Prvi od interfejsa je napajanje i hlađenje sa softverom za upravljanje serverskim resursima, koji se nalaze u slotovima/ketridžima, zbog čega je moguće izvršiti modelovanje upotrebe i raspodele opterećenja za jednog ili više korisnika.

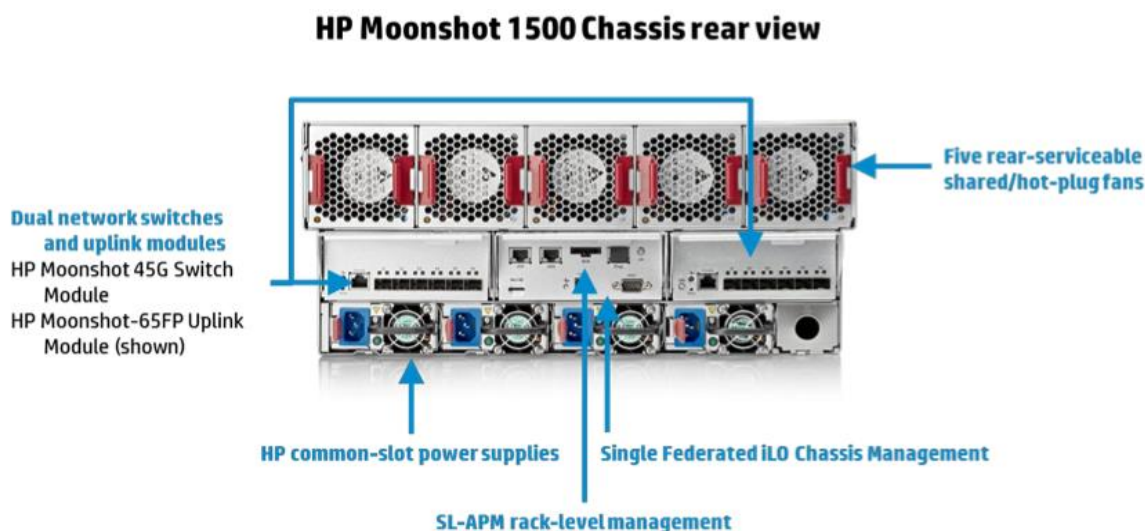
Drugi najvažniji deo za sisteme visokih performansi, je postojanost komunikacije sa tri interfejsa: **mreža, skladište podataka i klaster**. Svaki od ProLiant servera ima nezavisan pristup interfejsima. Neometana komunikacija unutar kućišta se realizuje pomoću dva switch modula koji podržavaju dva radna moda, prvi mod podrazumeva **ravnomernu** upotrebu propusnosti podataka (eng. Bandwith), dok drugi mod podrazumeva **maksimalnu**

propusnost podataka. Inicijalni switch modul sadrži implementaciju **1Gbps** linka, tako da svaki od servera može da ima **4x1Gbps** linka na svakom od switch modula, što je ukupno **8x1Gbps** linkova za komunikaciju unutar samog Moonshot kućišta.

Glavne karakteristike HP Moonshot 1500:

1. Unutar **10 Monshot servera** može se postaviti do **450 servera**, unutar standardnog 47U (širina 800mm, dubina 960mm, visina 2200mm) serverskog rek ormana (eng. Rack).
2. Kućište poseduje centralizovan sistem hlađenja i napajanja servera, sa softverom za menadžment i pojedinačno podešavanje svakog od servera.
3. Koristi **Intel Atom S1260** procesor.
4. Integrisana dva switch modula za direktnu komunikaciju između servera.
5. Podržava Linux, Windows i ostale operativne sisteme koji poseduju podršku za HPC.
6. Cena **\$61,875** za kućište sa **45 servera** i integrisanim switch-em.
7. Cena jednog servera bez serverskih modula je **\$1375**. [12]

Na prikazu (Slika 3.) se nalazi kućište sa switch modulima i slotovi za postavljanje HP ProLiant m710 servera, dok na prikazu (Slika 4.) se vide, različiti moduli za napajanje, računarsku mrežu, direktnu komunikaciju, serijski port itd.



Slika 4. Zadnji deo Moonshot 1500 kućišta

Izvor: http://www.citrix.de/content/citrix/en_us/articles-and-insights/trends-and-innovation/feb-2014/xendesktop-and-the-evolution-of-hardware-assisted-server-technologies/_jcr_content/par/imageframe_2/image2.img.png/1391722630086.png

Tabela 2. Hardverske karakteristike - Moonshot 1500 kućišta

CPU	Intel® Atom™ Processor S1260, 2.0 GHz
Memory Protection	Advanced ECC (multi-bit error protection)
Memory	Unbuffered (UDIMM), ECC @ 1333 MHz 1 SO-DIMM Slot 8GB (1x8GB)
Network	Network Controller Broadcom 5720 Dual-port 1Gb Ethernet Controller
Storage Controller	Marvell 9125
Storage Device	Hard Drive, Each cartridge will ship with a SATA HDD or SATA SSD
Boot Maximum Internal Storage (per Moonshot server)	Quick Release SATA 2.5" 1TB - 1 x 1TB SATA 500 GB - 1 x 500 GB SATA
Interfaces	Network RJ-45 for Management Two RJ-45 are located on the HP Moonshot 1500 Chassis Management Module. One is for the HP Moonshot iLO Chassis Management and one is for future use SFP+ connector Six 10GbE SFP+ connectors per HP Moonshot-6SFP Uplink. NOTE: SFP+ connectors are not included as standard option. NOTE: See the HP Moonshot-6SFP Uplink Module QuickSpecs for Technical Specifications and additional information: http://h18000.www1.hp.com/products/quickspecs/14734_na/14734_na.html Health LED 47 front for server/switch health, one per server and switch when access panel is removed Power 1 front and one per server and switch when access panel is removed UID for node 1 front and 1 rear (on management module) and one per server and switch when access panel is removed Drive health LED One per server and switch when access panel is removed
Industry Standard Compliance	ACPI 2.0b Compliant SMBIOS 2.6.1 PXE Support
Operating Systems and Virtualization Software Support for ProLiant	Canonical Ubuntu Red Hat Enterprise Linux (RHEL) SUSE Linux Enterprise Server (SLES)

Izvor: [9]

2.1.1.2 HP ProLiant m710

HP ProLiant m710 serverski modul specijalno izrađen i optimizovan **za potrebe obrade video materijala**, zasnovan na Intel® Xeon® E3-1284L v3 (1.8-3.2GHz/8-core/6MB/47W), radi na baznoj frekvenciji od 1.8 GHz ostvarujući maksimalnu brzinu 3.2 GHz upotrebom turbo boost tehnologije. ProLiant m710 ima integrisanu grafičku karticu Intel® Iris™ Pro 5200 DDR3 ECC 1.600 MHz 128MB sa 40 GPU jezgara koji zajedno rade na 1 GHz. Svaki od modula sadrži 2x10Gbps Mellanox mrežne kartice, koje se povezuju sa unutrašnjim switch modulom unutar Moonshot kućišta. Izbor za skladištenje podataka je NGFF SSD od 120 GB do 480 GB, koji služi za instalaciju operativnog sistema i skladištenje podataka. Potrošnja električne energije je od **30-60 W** u zavisnosti od opterećenja modula. [13] [14]



Slika 5. HP ProLiant m710 server

Izvor: <http://product-images.www8-hp.com/digmedialib/prodimg/lowres/c04460333.png>

Tabela 3. Hardverske karakteristike – HP ProLiant m710

CPU	Intel® Xeon® E3-1284L v3 (1.8-3.2GHz/8-core/6MB/47W)
Memory	32 GB (4) DIMMs 1,600 MHz, DDR3 SO-DIMM ECC (8 GB)
Network	Integrated NIC: dual port 10GbE Mellanox CX3 Supported switches: 45-port 10 Gb downlinks, (4) 40GbE QSFP uplinks
Storage	NGFF 120 GB and 480 GB SSD
Boot	Network or local via USB mezzanine
System configurations	• 45 m710 • 30 m710 • 15 m710
Power	Cartridge: < 60 W average
Form factor	Singlewide
OS	Ubuntu, RHEL, SLES, Microsoft® Windows® 2012

Izvor: [13]

2.1.2 HP Apollo 8000 sistem

Postavljeni izazov od strane tržišta za proizvođače u culju dobijanja većeg broja teraflopsa sa efikasnom potrošnjom električne energije, izražene u FLOPS/W sa povećanim brojem servera na jednom kvadratnom metru, što bi za ostvarivanje izazova upotrebom trenutne tehnologije, bilo bi potrebno oko 1 GW (gigawatt) električne energije, i 30-200 fudbalskih igrališta za smeštaj servera, kako bi se ostvario veći broj teraflopsa. HP kao odgovor na trenutne zahteve je predstavio prvo razvojno rešenje, koje koristi sistem vodenog hlađenja i gustog smeštaja servera, tako da unutar jednog rack-a može da se smesti 72 HP ProLiant XL730f servera. [15]

HP Apollo 8000 sistem je dostupan u **razvojnoj verziji** sa skalabilnom početnom konfiguracijom, od jednog **HP Apollo f8000 rack-a** i jedne **HP Apollo 8000 iCDU** rack jedinice, pri čemu sistem može da sadrži do 144 servera unutar 2 x HP Apollo f8000 rack-a. Sistem poseduje dosta akceleratora, PCIe kartica, propusnog opsega, zaštitna podešavanja koja ispunjavaju standarde, koji se zahtevaju od strane tržišta i budućih korisnika ovog sistema. Karakteristike po kojima se izdvajaju **HP HPC** sistemi je u ostvarenim teraflopsima po reku i njihovim rezultatima. Efikasnost vodenog hlađenja donosi održivost budućih data centara, osiguravajući konstantnost performansi i hlađenja.

HP Apollo 8000 se sastoji od nekoliko ključnih komponenti:

1. HP Apollo f8000 rack, unutar kojeg se smeštaju HP ProLiant XL730f serveri.
2. HP Apollo 8000 iCDU Rack, centralna jedinica za vodeno hlađenje koja podržava do 4 x HP Apollo f8000 rack-a.
3. HP ProLiant XL730f server
4. HP Infiniband switch za HP Apollo 8000 sistem, switch zadužen za komunikaciju između HP ProLiant XL730f servera.



Slika 6. Sklopljen Apollo sistem

Izvor: [16]

2.1.2.1 HP Apollo f8000 rack

U okviru osnovne konfiguracije f8000 postoji podrška za smeštanjem do 72 servera (HP ProLiant XL730f). Rack poseduje mogućnost ugradnje do **8 HP InfiniBand switch-eva**, koji dolaze sa 16 SFP+ kablova. Centralizovani sistem za napajanje, smanjuje mogućnost otkazivanja sistema. Sistem se napaja sa ulaznim napajanjem od 380-480VAC zbog čega se prilikom snabdevanja i distribucije do servera, vrši konvertovanje i prilagođavanje napona. Blok za napajanje poseduje mogućnost napajanja sa jednog izvora (N+1), više izvora (N+N), ili iz više rezervnih izvora (N+1, N+N) električne energije (Slika 7.). Distribucijom električne energije upotrebom centralizovanog sistema za napajanje se ostvaruje efikasnost čak i do 90%.

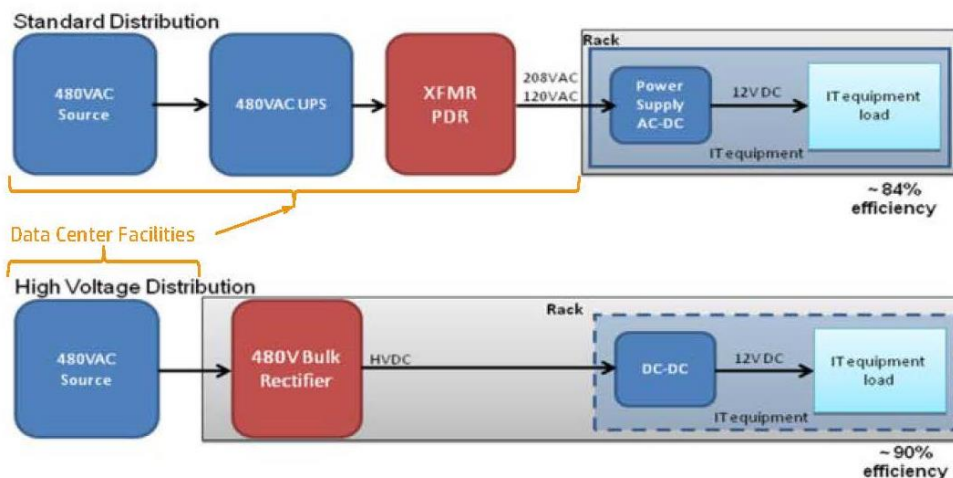
Tabela 4. Karakteristike HP Apollo f8000 rack-a

Server	Each rack supports up to 144 (1 tray – 72 servers) HP ProLiant XL730f Server trays (2 nodes per tray)
Server Networking	Each rack supports a total of 8 HP InfiniBand switches
Power	80 kW input power per rack ships standard with N+1 or N+N redundancy support depending on configuration of the servers
Input:	380-415VAC for international standards and 480VAC for North American standards (4x30A power cords per rack)
Management	HP Apollo 8000 System Manager HP iLO Management Engine (iLO 4 v2.00) Rack level iLO network consolidation
Typical configuration	72 HP ProLiant XL730f Server trays and 8 HP InfiniBand switches, 16 Ethernet SFP+ cable kits, associated rack plumbing kit and utility module (includes HP Apollo 8000 System Manager, 2x40KW power shelves)
Weight	4700 lbs (or 2132kg) Max 2914 lbs (or 1322kg) Max with no server trays
Dimensions (WxDxH)	24 in x 56.18 in x 94 in (607mm or 1427mm or 2382mm)

Izvor: [16]

High-Voltage AC to the Rack: Limiting conversion steps to improve efficiency

Power System Options



Slika 7. HP Apollo 8000, blok za snabdevanje i konverziju električne energije

Izvor: [17]



Slika 8. HP Apollo f8000 rack

Izvor: <http://product-images.www8-hp.com/digmedialib/prodimg/lowres/c04326754.png>

2.1.2.2 HP Apollo 8000 iCDU Rack

Hlađenje kao jedna od ključnih stavki za funkcionisanje servera i drugih komponenti koje učestvuju u HPC sistemu, zahteva upotrebu određenog načina hlađenja. Specifičnost HP Apollo sistema je u upotrebi centralizovanog vodenog hlađenja sa CDU sistemom zatvorenih cevi, ventila, pumpi;. Koristeći hladnu vodu za hlađenje, i toplu vodu za grejanje prostorija (Slika 10.). Voda kao medijum za hlađenje, ima oko 1.000 puta veću toplotnu efikasnost, naspram hlađenja vazduhom, zbog čega jedna čaša hladne vode ima kapacitet kao jedna soba hlađena vazduhom, pri čemu potrebna energija za pumpu koja će vršiti protok vode zajedno sa distribucijom toplote do grejnih tela je znatno manja, naspram sistema hlađenja vazduhom.

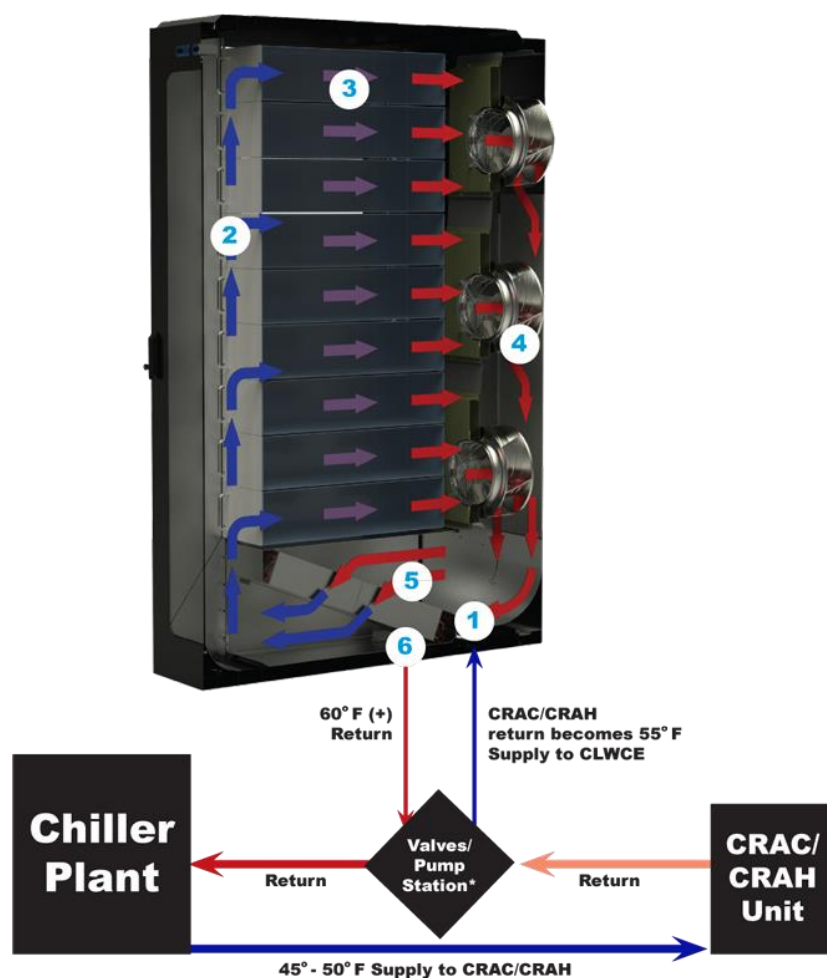
HP Apollo koristi zatvoreni sistem toplotnih cevi, koje se koriste za cirkulaciju vode upotrebom CDU (eng. CDU - coolant distribution unit) jedinice, koja je spojena zajedno sa HVAC sistemom (eng. HVAC - **h**eating, **v**entilating, and **a**ir **c**onditioning) za distribuciju električne energije unutar HP Apollo sistema. CDU predstavlja jedinicu za upravljanje protokom (tople i hladne) vode. Održavajući konstantnu temperaturu **tople** vode ~65°F (18.33°C) i **hladne** vode ~45°F (7.22°C) (Slika 10,11.).

Upotrebom vodenog hlađenja, vrši se direktan transfer toplote sa procesora i svih komponenti unutar kućišta, smanjujući troškove čak i do \$1.000.000 po godini, pružajući mogućnost uštede i ulaganja u dalji razvoj kompanije i njenih zaposlenih. [16] [17] [18]



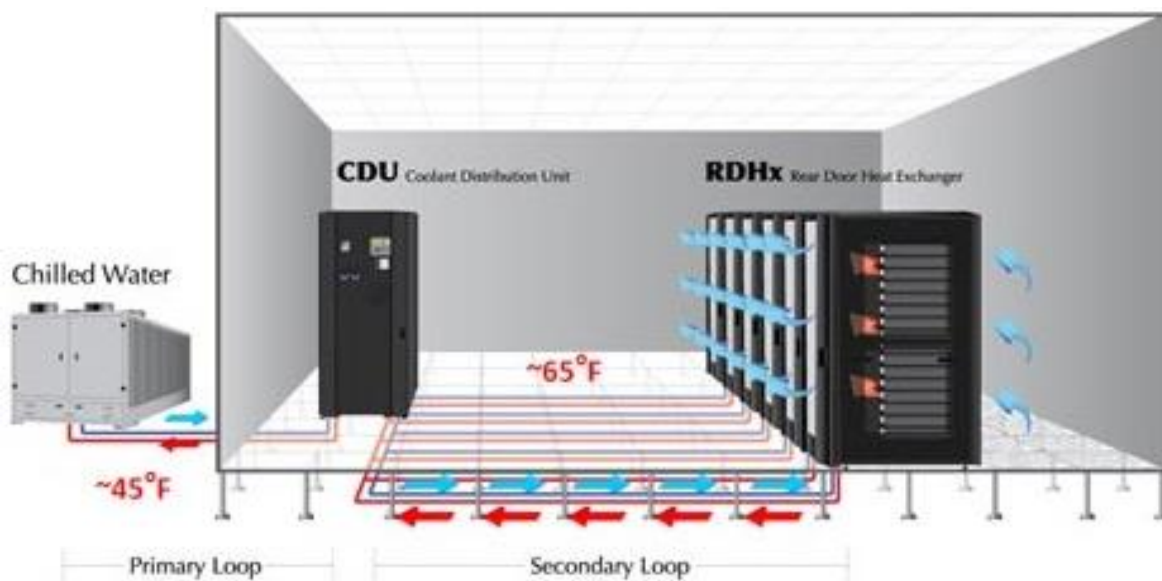
Slika 9. HP Apollo 8000 iCDU Rack

Izvor: <http://product-images.www8-hp.com/digmedialib/prodimg/lowres/c04317164.png>



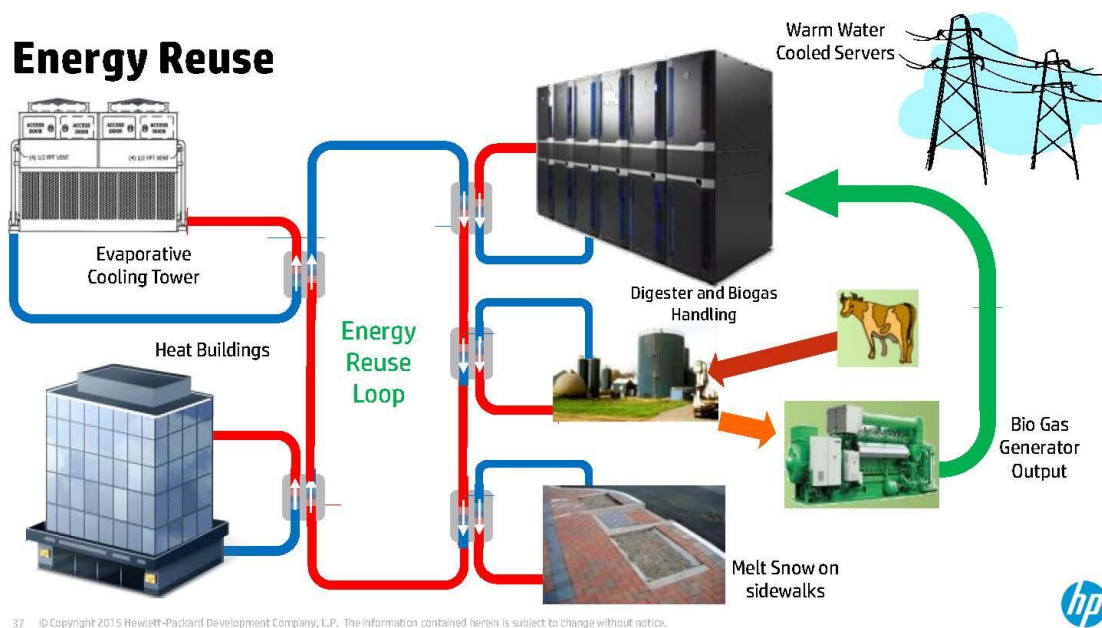
Slika 10. CDU jedinicu za upravljanje protokom tople i hladne vode.

Izvor: http://international.werackyourworld.com/Solutions/cooling/img/CLWCE_Flow11.png



Slika 11. Sistem vodenog hlađenja sa CDU jedinicom.

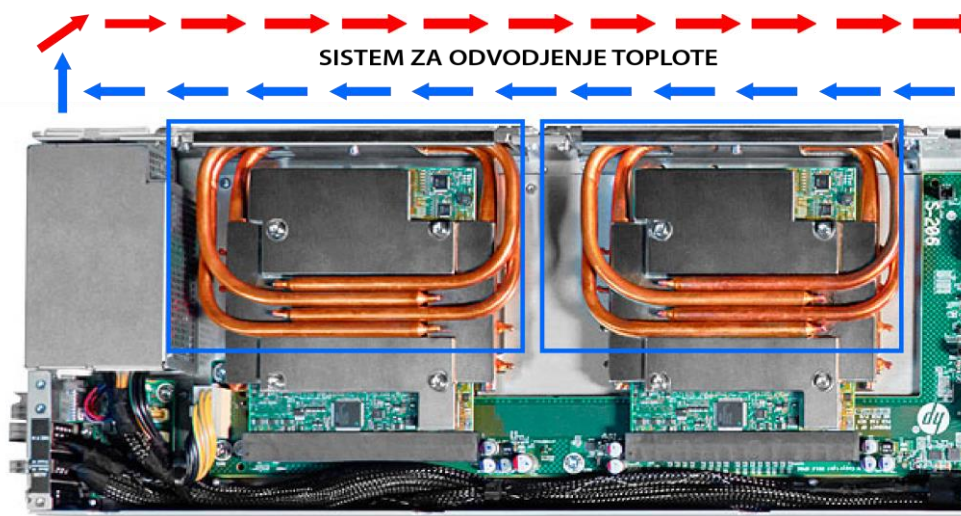
Izvor: <http://www.datacenterknowledge.com/wp-content/uploads/2010/02/coolcentric-cdu.jpg>



Slika 12. Primer funkcionisanja CDU sistema, unutar jednog data centra.

Izvor: [17]

Na prikazu (Slika 13.) se nalazi sistem odvođenja toplote sa jednog od servera, upotrebom centralne CDU jedinice. Za odvođenje toplote se koriste dve bakarne cevi, pri čemu jedna (plava strelica) vrši hlađenje servera, a druga odvođenje toplote (crvena strelica) pomoću CDU jedinice.



Slika 13. Primer, sistem za odvođenje toplote sa jednog servera.

Izvor: <http://spectrum.ieee.org/img/hp%20innards%20blog-1402409874193.jpg>

Zahvaljujući HP serverskoj tehnologiji suvog uklanjanja (eng. dry-disconnect), komponente su uvek suve i očuvane na određenoj temperaturi. Razlog za to je inovacija u distribuciji električne energije upotrebom iCDU (eng. intelligent Cooling Distribution Unit) rek tehnologije i vodenog hlađenja, koje zadovoljava ENERGY STAR Platinum serifikaciju. Zbog efikasnosti hlađenja, moguće je postaviti više servera u manjem prostoru. HP Apollo sistem pruža mogućnost **direktne ekstrakcije** toplote sa procesora i komponenti, što omogućava postizanje većih brzina i performansi, koje u suprotnom nebi bile moguće. Apollo sistem može da sadrži od 72 servera pa do nekoliko 1.000, pri čemu jedan server ima potrošnju od ~500W, dok celokupan sistem ima potrošnju od **80 kW** (4 x 30A 3ph 480AC). [16] [17] [18] Za napajanje električnom energijom se koristi industrijski priključak i specijalizovani konektori zajedno sa sklopkama, koje mogu da izdrže opterećenje od **80 kW**.

Karakteristike **HP Apollo 8000 iCDU** rack-a:

1. Gust raspored, na vrhu rack ormana mogućnost za smeštaj dodatnih uređaja, unutar 26U razpoloživog prostora.
2. Podržava hlađenje do **4 HP Apollo f8000 rack-a**.
3. Ne zahteva upotrebu specijalne vrste vode, moguća je upotreba bilo kog tečnog oblika vode.
4. Izbegava se ograničenje i upotreba vazdušnog hlađenja.
5. Smanjuje potrošnju i povećava uštedu el. energije.
6. Posедуje vakum pumu koja održava konstantan pritisak, sprečavajući isticanje vode i održavajući je na mestu usled pucanja cevi.
7. Brzo, lako instaliranje i povezivanje sa sistemom za vodo snabdevanje.
8. Tehnologije tri petlje, održava vodeni pritisak izolovajući snabdevanje vodom unutar objekata i drugu petlju vodo snabdevanja (Slika 12.). [16]

2.1.2.3 HP ProLiant XL730f server

Upotrebom HP ProLiant XL730f Gen9 servera unapređuje se proces naučnih otkrića, zbog veličine i broja servera (144 servera) unutar dva f8000 rack ormara. HP ProLiant XL790f Gen9 pokreće Intel® Xeon® koji obezbeđuje potrebnu računarsku snagu za rukovanjem različitim zahtevnim zadacima. Podržava do 256 GB RAM memorije po nodu, sa opcionim SSD hard diskom. Posедуje **1x FDR InfiniBand** port sa **1Gbps NIC** mrežnom konekcijom, poboljšavajući povezanost i dostupnost podataka. Potpuno novi ProLiant XL730f Gen9 sa HP ProActive Insight arhitekturom pruža još veći nivo za konfiguraciju super računarskog okruženja, koji zahtevaju najvišu moguću gustinu performansi i efikasnosti hlađenja, sa usmerenjem ka reciklaži energije. Tehnologija isključivanja na suvo (Dry disconnect) obezbeđuje dovoljno rashladne tečnosti preko sistema zatvorenih bakarnih cevi, razmenjujući toplotu preko termalnih sabirnica u reku. Ostvarivajući obećanje životnog ciklusa automatizacije, ProLiant XL730f Gen9 poseduje ekskluzivne mogućnosti HP Active Health sistema koji obezbeđuje kontinuirano, proaktivno, praćenje preko 1600 sistema bez uticaja na performanse aplikacija koje se trenutno izvršavaju na serveru. [16]



Slika 14. HP ProLiant XL730f server

Izvor: <http://product-images.www8-hp.com/digmedialib/prodimg/lowres/c04317309.png>

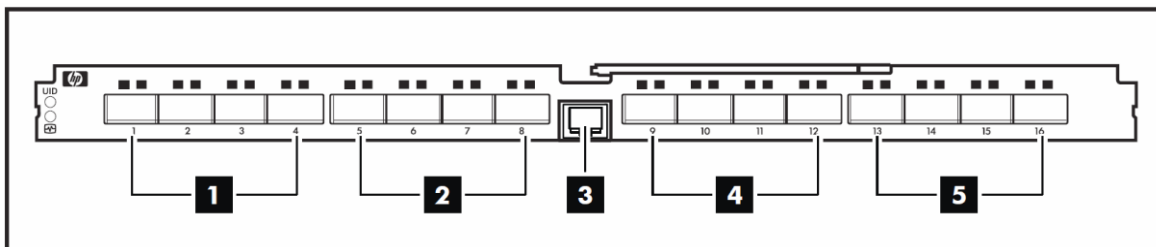
Tabela 5. Hardverske karakteristike – HP ProLiant XL730f

CPU	Intel® Xeon® processor E5-2600 v3
Memory	16 DIMMS total per server or max 256GB memory per server DDR4 support on future Intel® Xeon® processor ES-2600 v3 product family
Network	Integrated NIC: Single port 1GbE per server IB Adaptor Kit: Single ConnectX-3 Pro FDR InfiniBand port per server
Storage	1 Small form factor SSD per server Supports 80GB/120GB SSD
Boot	SSD and network
System configurations	2 CPUs per server, single FDR InfiniBand adaptor, 2 DIMMs per CPU (up to 8 DIMMS max)
Power	Max of 1200W of HVDC to 12V conversion per ProLiant XL730f Server tray
Form factor	1U half width
Managment	• HP Apollo 8000 System Manager • HP iLO Management Engine (iLO 4) - dedicated iLO network support • HP Advanced Power Manager • HP Insight Cluster Management Utility
OS	RHEL, SLES, CentOS

Izvor: [16]

2.1.2.4 HP Infiniband switch

Slika 15 prikazuje Quad Small Form-factor Pluggable (QSFP) portove za konektovanje na prednjoj strani HP BLc 4X QDR InfiniBand Switch-a. Infiniband switch koristi više različitih sistema za komunikaciju radi ostvarivanja određenih brzina, koje su potrebne HPC sistemu.



Slika 15. HP BLc 4X QDR InfiniBand Switch Connector Bays

Izvor: [19]

Na prikazanoj slici (Slika 15.) se nalazi opis komponenti:

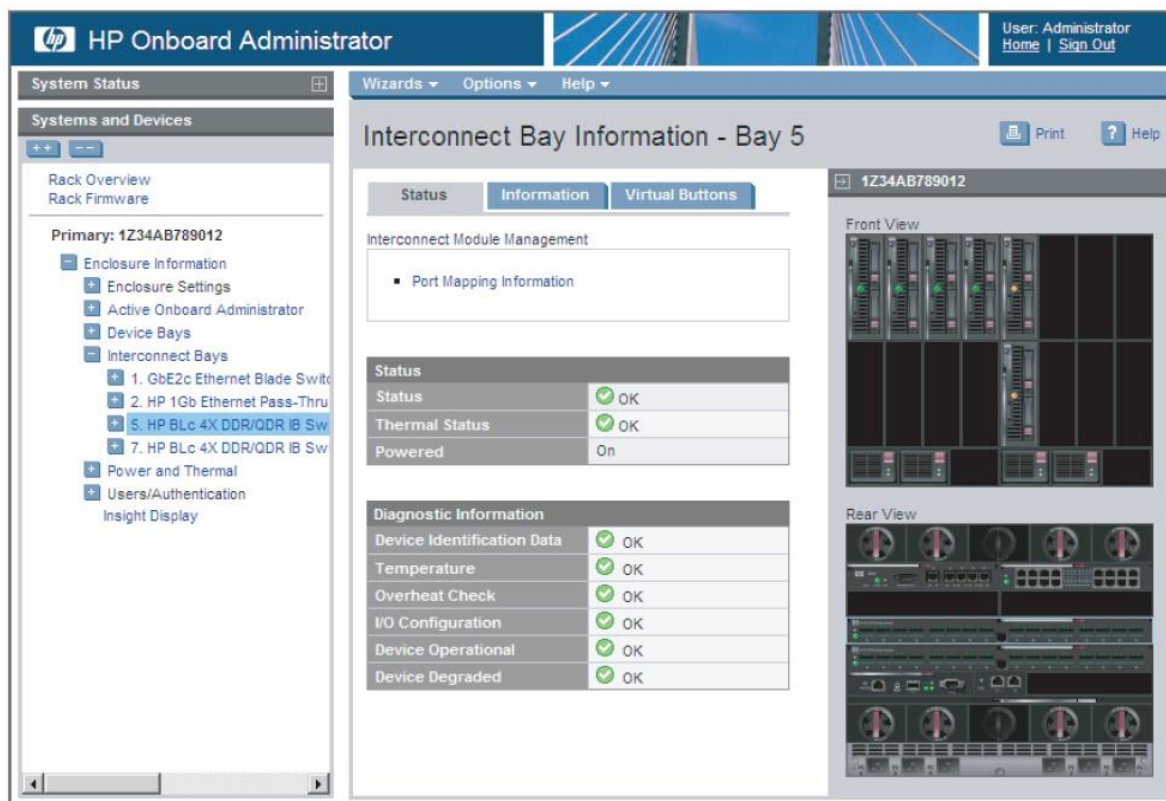
1. Uplink konektori, portovi 1-4
2. Uplink konektori, portovi 5-8
3. I²C port (inter-integrated circuit port)
4. Uplink konektori, portovi 9-12
5. Uplink konektori, portovi 13-16

Tabela 6. Karakteristike HP Infiniband switch-a

Performance and Form Factor	
Switch Blade Type	Double wide switch blades (switch bays 5&6, 7&8)
Performance	40Gbps per port for 4X QDR, and 20Gbps per port for 4X DDR
Port Configuration	The HP BLc 4X QDR IB Switch • 16 internal 4X QDR downlink ports • 16 external 4X QDR QSFP uplink ports The HP BLc 4X DDR IB Gen 2 Switch • 16 Internal 4X DDR downlink ports • 16 external 4X DDR QSFP uplink ports The QLogic BLc 4X QDR IB Switch • 16 Internal 4X QDR downlink ports • 16 external 4X QDR QSFP uplink ports
Media Types	QSFP/CX4 copper, QSFP copper, QSFP fiber optic cables
Management and Protocols	
Management Features	Embedded subnet manager on rack mount switches, or host-based subnet manager, or management module for QLogic switch blade.
Software Supported	OFED for Linux and WinOF for Windows.
Deployment	
Max per Enclosure	Switch blades: maximum 2. Refer to configuration documents for more details.
Options Available	Switches: 489184-B21: HP BLc 4X QDR IB Switch for HP BladeSystem c-Class 489183-B21: HP BLc 4X DDR IB Gen 2 Switch for HP BladeSystem c-Class 505958-B21 QLogic BLc 4X QDR IB Switch for HP BladeSystem c-Class 505959-B21 QLogic BLc 4X QDR IB Management Module Mezzanine HCAs: 592519-B21: HP 4X QDR IB CX-2 Dual-port Mezzanine HCA for HP BladeSystem c-Class 448262-B21: HP 4X DDR IB Dual-port Mezzanine HCA for HP BladeSystem c-Class 583210-B21 QLogic 4X QDR IB Dual Port Mezz HCA for HP BladeSystem c-Class

Izvor: [20]

HP Infiniband switch sadrži pametan sistem (eng. Smart Sistem) za administraciju (Slika 16.), na kojem su moguća podešavanja rutiranja, ip adrese, brzina povezivanja itd. Takođe može se koristiti i posebno kao centralizovani Switch za sisteme čuvanja podataka, ili komunikaciju između radnih stanica.



Slika 16. HP Infiniband switch Administracija.

Izvor: <http://h10032.www1.hp.com/ctg/Manual/c01700535>

Glavne karakteristike HP Apollo 8000 sistema

1. Uklanjanje servera upotrebom tehnologije, uklanjanja na suvo (eng. Dry-disconnect), održavajući komponente suvima i hlađenima na određenoj temperaturi.
2. HVAC sistem za distribuciju električne energije, **eliminise** standardne konverzacione korake koji **prelaze** Energy Star Platinum sertifikaciju, ostvarujući manju potrošnju električne energije.
3. HP Apollo 8000 sa svojim iCDU rack-om održava pod atmosferskim pritiskom operativnu izdržljivost celokupnog sistema i njegovog servisiranja.
4. Brzina dodavanja modula sa posedovanjem modularnog kompleta za vodeno hlađenje, koje se nalazi unutar svakog od servera zbog čega serveri mogu biti operativni u toku nekoli sati za razliku od drugih sistema.

2.1.3 Uporedni prikaz karakteristika HP

Tabela 7. Uporedni prikaz karakteristika HP servera

Server Name	HP ProLiant m710	HP ProLiant XL730f Server
CPU	Intel® Xeon® E3-1284L v3(1.8-3.2GHz/8-core/6MB/47W)	Intel® Xeon® E5-2600 v3
Memory	32 GB (4) DIMMs 1,600 MHz, DDR3 SO-DIMM ECC (8 GB)	16 DIMMS total per server or max 256GB memory per server DDR4 support on future Intel® Xeon® processor E5-2600 v3
Network	Integrated NIC: dual port 10GbE Mellanox CX3 Supported switches: 45-port 10 Gb downlinks, (4) 40GbE QSFP uplinks	Integrated NIC: Single port 1GbE per server IB Adaptor Kit: Single ConnectX-3 Pro FDR InfiniBand port per server
Storage	NGFF 120 GB and 480 GB SSD	1 Small form factor SSD per server Supports 80GB/120GB SSD
Boot	Network or local via USB mezzanine	SSD and network
System configurations	• 45 m710 • 30 m710 • 15 m710	2 CPUs per server, single FDR InfiniBand adaptor, 2 DIMMs per CPU (up to 8 DIMMS max)
Power	Cartridge: < 60 W average	1 Server ~500W Max of 1200W of HVDC to 12V conversion per ProLiant XL730f Server tray
Form factor	Singlewide	1U half width
OS	Ubuntu, RHEL, SLES, Microsoft® Windows® 2012	RHEL, SLES, CentOS
Number of servers	45	72
Single server	\$1375	Not available
Price System	\$61,875	Not available

Izvor: sopstvena izrada, dobijeni podaci na osnovu istraživanja unutar rada.

HP ProLiant XL730f server trenutno nije dostupan na tržištu, jer još uvek se nalazi u procesu laboratorijskog testiranja i izgradnje, zbog čega još uvek nije poznata cena krajnjeg proizvoda.

2.2 IBM

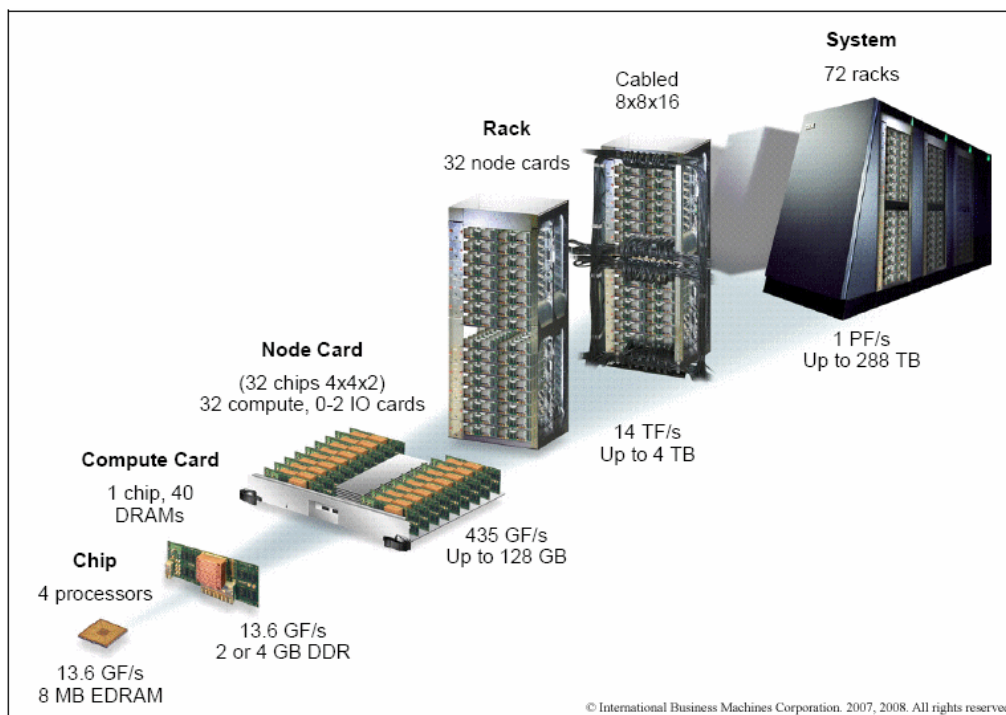
IBM (International Business Machines) je kompanija sa najdužom tradicijom u oblasti informacionih tehnologija, sa sedištem u New Yorku. Kompanija je specijalizovana za proizvodnju hardvera, softvera, servera, skladišta podataka, mikročipova, procesora i specijalizovanog aplikativnog softvera. Dominantni položaj na tržištu, IBM je ostvario zahvaljujući ulaganjima u istraživanja, inovacije i proizvodnju, izdvajajući na godišnjem nivou i do 50% ukupnog budžeta za potrebe istraživanja, konstantno impresionirajući investitore i istraživače sa svojom sposobnošću osluškivanja i prilagođavanja proizvoda potrebama tržišta. IBM u svojoj ponudi sistema visokih performansi poseduje dva sistema: **IBM Blue Gene/Q** i **IBM Power 775**.

2.2.1 IBM Blue Gene

Blue Gene je IBM-ov projekat sa modularnim dizajnom super računara koji mogu da dostignu operacione brzine izražene u **PFLOPS** (petaFLOPS) opsegu, sa manjom potrošnjom energije. Takođe koristi tehnologiju vodenog hlađenja i sistema cevi, kao i njegovi konkurenti.

IBM Blue Gene/Q trenutno predstavlja najefikasniji i prilagodljiviji sistem za intenzivne računske operacije, namenjen opštim naučnim aplikacijama, čineći ga najbržim analitičkim sistemom na planeti, koji se nalazi među prvih 10 HPC sistema na Graph500 listi [21]. Jedan je od lidera u energetskej efikasnosti među najboljim zelenim sistemima (eng. Green Systems) objavljenima na Green500 listi. Blue Gene/Q pripada porodici modularnih super računara, koji su prepoznatljivi po efikasnosti u napajanju, hlađenju i racionalnoj upotrebi fizičkog prostora. Predstavlja idealan sistem za rad sa velikim skupovima podataka, njihovom analizom, renderovanjem grafikona, modeliranjem, istraživanjem itd.

Celokupan sistem je sastavljen iz više modula (Slika 17.) koji se kasnije spajaju u jednu celinu pod nazivom **JuQueen**. Koja ima ulogu upravlja celokupnim sistemom raspoređujući radne zadatke i direktive vezane za procesiranje podataka. [22] [23] [24]



Slika 17. Sistem IBM Blue Gene/Q.

Izvor: <https://computing.llnl.gov/tutorials/bgp/images/bgpScalingArch.gif>

Niz tehničkih inovacija u hardveru, softveru i modularnom dizajnu sistema pruža Blue Gene korisnicima postizanje novih nivoa performansi i energetske efikasnosti ostvarujući brzinu i propusnu moć nekoliko hiljada ili miliona računara. Blue Gene čini niz različitih modula, posедуjući fleksibilnost modelovanja prilikom kupovine, jer je moguće projektovanje potrebne računarske snage, na osnovu čega se određuje broj potrebnih rack-ova i računskih jedinica (eng. Compute node).

Na primer:

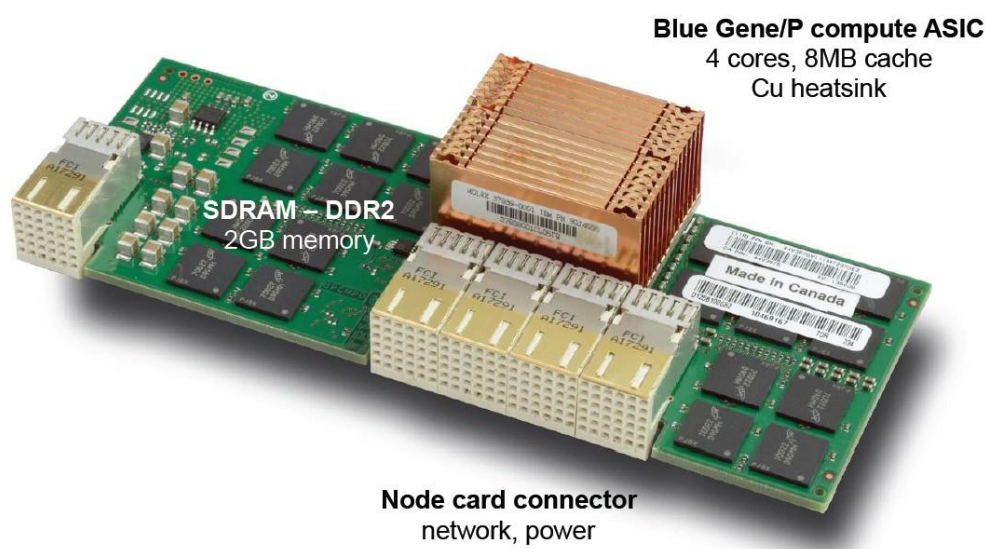
1. 8 rack-ova = 1.7 PF/s
2. 28 rack = 5.9 PF/s
3. 96 rack-ova = 20 PF/s

Osnovna svakog sistema je procesor i njegova mogućnost izračunavanja i obrade određene količine informacija. U ovo slučaju procesor je modularnog tipa, tako da Blue Gene koristi modularne računске kartice (Slika 18.) koje sadrže:

1. 4 jezgra sa 8 MB cache memorije.
2. SDRAM – DDR2 2GB ili SDRAM – DDR3 16GB.
3. Node kartica sa konektorom (Node card connector).
4. Računarska mreža (Network).
5. Napajanje (Power).
6. Sistem za vodeno hlađenje CU heatsink.

Svaka od kartica se smešta unutar Blue Gene/P node card kućišta koji povezuje sve kartice u jednu logičku celinu (Slika 19.), tako da svaka Blue Gene/P računска kartica (eng. Compute card) poseduje napajanje, centralizovano hlađenje i mrežnu komunikaciju. [25]

Blue Gene/P compute card

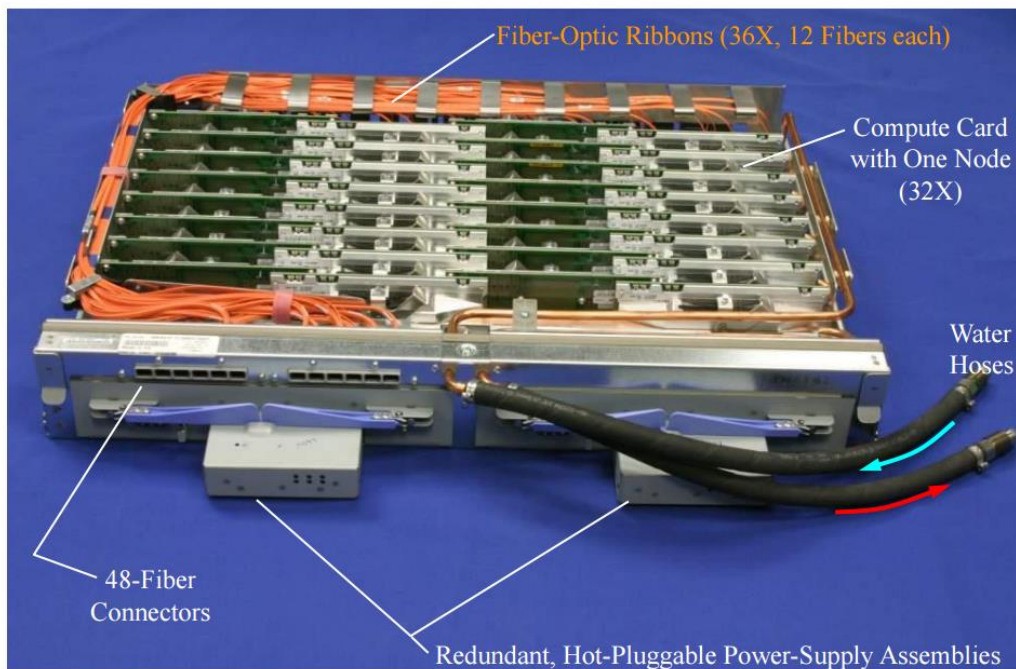


Slika 18. Blue Gene/P compute card sa 4 jezgra i 8 MB cache memorije.

Izvor: [23]

Blue Gene/P node card kućište poseduje podršku za smeštanjem do 32 Blue Gene/P compute card modula unutar **dva kućišta** (2x2x2x2). Na svakom od kućišta se nalazi određeni broj zatvorenih cevi, koje se koriste za potrebe vodenog hlađenja (Slika 19,20.). Povezivanje između nodova se vrši upotrebom **36 optičkih kablova**, svaki optički kabl poseduje 12 optičkih vlakana koji se kasnije zajedno mogu povezati na računarsku mrežu upotrebom **48 Fiber Optic** konekcija.

Blue Gene/Q node card



Source: IBM

Slika 19. Blue Gene/Q node card, Primer sastavljenog noda.

Izvor: [23]

Blue Gene/P node card



Slika 20. Blue Gene/P node card kućište sa compute modulima.

Izvor: [23]

Blue Gene/Q sistem se sastoji iz 3 glavna dela:

1. Interaktivnih nodova (front-end ili nodovi za logovanje korisnika)
2. Servisnih nodova
3. Fizičkih uređaja za izračunavanje, procesori (Power 7)

Front-end predstavlja ulaznu tačku sistema gde korisnik ima direktan pristup dodeli zadataka i njihovoj kompilaciji. Front-end mašina se pokreće pod Linux operativnim sistemom. Spajanje 32 jezgra procesora Power 7, dostiže se brzine od ~3GHz i 64 GiB (~68.71 GB) pri čemu krajni korisnik vidi celokupan sistem kao jedan računar. [25] [26]

Servisni nodovi su zaduženi za upravljanje resursima (zadacima, bazama podataka, rasporedom opterećenja itd.)

Fizički uređaji koji sačinjavaju Blue Gene sistem, počinju i završavaju sa konfiguracijom osnovnih elemenata:

1. 6xBlue Gene/Q rack ormana.
2. 1x rack sa 32 računskih nodova (eng. compute nodes).
3. Svaki rack podeljen na 2 celine (eng. midplanes).
4. Svaki compute node sadrži 16 jezgri.
5. Svako jezgro poseduje podršku za pokretanjem do **4 procesa/niti** (eng. thread)

Svaki nod poseduje **16 GiB (~17.17 GB)** ram memorije i teoretsku snagu od **201.1 Gflop/s**, što je približno **12.7 Gflop/s** po jezgru. Tako celokupna sastavljena konfiguracija može da sadrži do **6.144** računarskih nodova sa **98.304** procesorskih jezgara i **393.216** hardverskih niti, ostvarujući kulminacionu snagu od **~1.258 Pflop/s** i **96 TiB (~105.55 TB)** ram memorije. Svaki od rack-ova poseduje **16** I/O nodova sa 2 konekcije po nodu, zadužene za komunikaciju sa računskim nodovima. [25] [26]

Računske nodove koji sačinjavaju **JuQueen (Slika 21.)**, pokreće lagani x64 Linux operativni sistem pod nazivom **CNK** (eng. Compute Node Kernel), koji rezerviše zadnje jezgro za izvršavanje svojih operacija. U cilju sprečavanja smetnji prilikom upotrebe ostalih procesorskih jezgara, koristi se prilagođeni minimalni kernel zadužen za osnovne servise: menadžmenta procesa, memorije, otklanjanje grešaka (eng. debugging), pristupa mreži, pristupa soketima (eng. sockets), dinamičkim bibliotekama, dostupnosti i podrškom za udaljenim RAS servisiranjem (eng. RAS – Remote Administration System). CNK poseduje većinu Linux sistemskih poziva, koji imaju ulogu prosleđivanja I/O operacija specijalizovanim I/O nodovima unutar sistema, koji kasnije izvršavaju različite operacije poput čitanja i pisanja podataka.

CNK poseduje podršku za različite komunikacione protokole kao što su MPI, OpenMPI, pthreads, CHARM+, ARMCI, UPC, PAMI i SPI. I/O nodovi imaju pristup eksternim uređajima, rutirajući zahteve koje dobijaju od CNK operativnog sistema.

Blue Gene/Q računski nodovi zajedno pokreću prilagođeni (eng. patched) **Red Hat Enterprise Linux 6**, optimizovan za potrebe rada na Blue Gene/Q HPC sistemu, sa glavnom namenom postojanosti podrške za više paralelnih file sistema kao što su: **GPFS, NSF, PVFS i Lustre**. [23] [24]

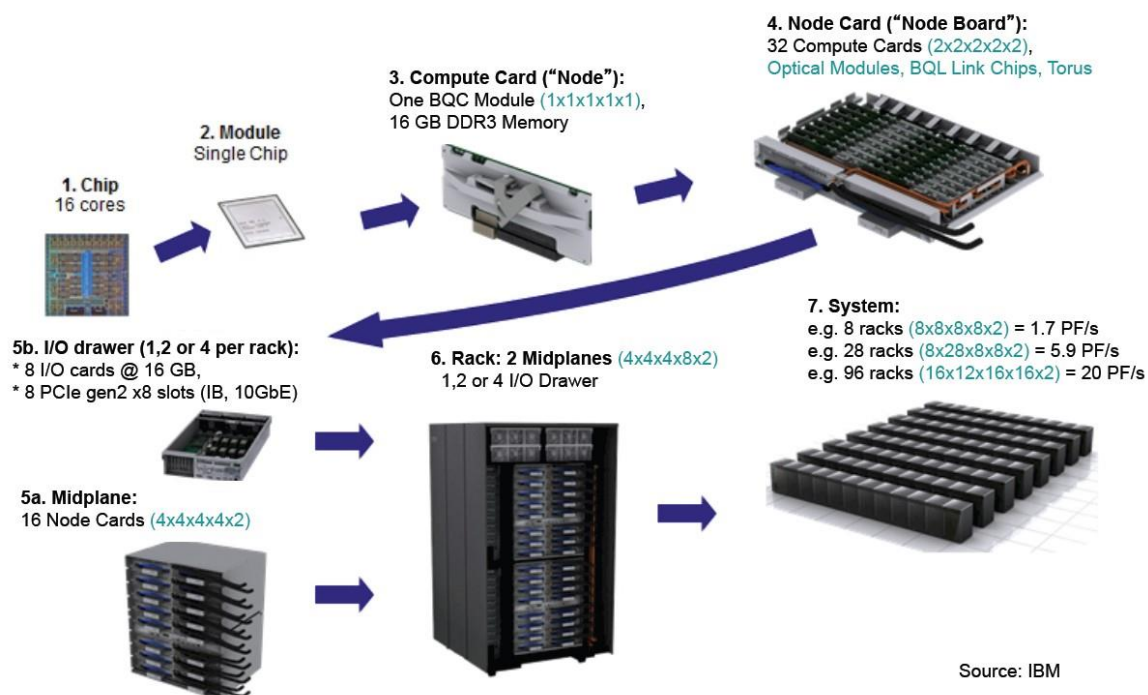
Front end nodovi ili nodovi za logovanje korisnika za svoj rad koriste originalnu verziju Red Hat Enterprise Linux 6 operativnog sistema pružajući korisnicima celokupno radno okruženje, koje može biti prilagođeno njihovim potrebama. [25] [26]

Sastavljanje svakog Blue Gene/Q sistema i stvaranje **JuQueen** arhitekture je sadržano od **7 koraka**, koji podrazumevaju sastavljanje određenih modula. U zavisnosti od broja sklopljenih nodova ostvaruje se određena brzina izraženu u PF/s.

Komponente koje sačinjavaju sklopljenu (Slika 21.) **JuQueen** arhitekturu:

1. 1x procesorski čip sa 16 jezgara
2. 1x Module, 1 čip
3. Compute kartica ("Node"), 1x BQC modul (1x1x1x1x1) 16 GB DDR3 ram memorija.
4. Node kartica ("Node Board"), 32 compute kartice (2x2x2x2x2), Optički moduli, BQL Link Chips, Torus.
5. Rack orman za kućišta sadrži:
 - a. 2 celine (Midplane) kućišta, 16 node kartica (4x4x4x2)
 - b. I/O ladlica, 8x I/O kartica sa 16GB, 8 PCI-E gen2 x8 slotovi (IB, 10GbE)
6. Rack: 2 Midplanes (4x4x4x8x2), 1,2 ili 4x I/O ladlice
7. Različite konfiguracije sistema:
 - a. 8 rack-ova (8x8x8x8x2) = 1.7 PF/s
 - b. 28 rack-ova (8x28x8x8x2) = 5.9 PF/s
 - c. 96 rack-ova (16x12x16x16x2) = 20 PF/s

Blue Gene/Q design



Slika 21. Blue Gene/Q Desing, proces sklapanja sistema.

Izvor: [23]

2.2.1.1 Analiza JuQueen konfiguracija

U zavisnosti od konfiguracije JuQueen arhitekture, ostvaruju se različite performanse što značajno utiče na potrošnju električne energije, potreban prostor, finansijsku isplativost. Analiziramo dve Blue Gene/Q konfiguracije, koje se sastoje od **8** i **28** rack ormana, pri čemu **8 rack-ova** ima potrošnju od **400 kW**, a **28 rack-ova** ima **1.4 MW**, što znači da bi konfiguracija od **96 rack** ormana imala potrošnju od **~4.76 MW**.

Tabela 8. Blue Gene/Q JUQUEEN konfiguracije

8 rack-ova	28 rack-ova
8.192 Compute Nodes (16 cores, 16 GB memory)	28.672 Compute Nodes (16 cores, 16 GB memory)
131072 cores / 0.5M treads	458752 cores / 1.8M threads
1.68 PFlop/s peak performance	5.88 PFlop/s peak performance
88 I/O nodes (10GigE) ← (1x32 + 7x8)	248 I/O nodes (10GigE) ← (1x32 + 27x8)
400 kW power consumption (10-50 kW per rack)	1.4 MW power consumption (10-50 kW per rack)
2 Frontend Nodes (user login) + 2 Service Nodes (system, database)	4 Frontend Nodes (user login) + 2 Service Nodes (system, database)
IBM p7 740, 8 cores (3.55 GHz), 128 GB memory	IBM p7 740, 8 cores (3.55 GHz), 128 GB memory
local storage device DS5020 (16 TB)	local storage device DS5020 (16 TB)

Izvor: [25]

Tabela 9. Karakteristike jednog Blue Gene/Q noda.

CPU	1.6 GHz 64-bit POWERPC A2, L1:16 KiB + L1d: 16 KiB, L2: 32 MiB
Memory	16 GiB DDR3
Storage	SAN, NAS
Boot	HDD or SSD
System configurations	Linux Batch commands
Power	55W
OS	Red Hat Enterprise Linux 6
Number of servers	8.192, 28.672, 98.304
Price	\$ 90.000 [27]

Izvor: sopstvena izrada, dobijeni podaci na osnovu istraživanja unutar rada.

2.2.2 IBM Power 775

IBM Power 775 kao i ostali IBM HPC sistemi, pruža kapacitet za pokretanje i izvršavanje različitih računskih algoritma. Zasnovan je na Power 7 procesorskom jezgru, dizajniranom za organizacije koje zahtevaju visoku skalabilnost sistema, sa ekstremnim paralelnim procesiranjem i performansama koje se mogu smestiti unutar gustog modularnog pakovanja. Upotrebljava se u grupisanim klasterisanim konfiguracijama sa najmanjim brojem od 512 procesorskih jezgara, ili u klasi svetskih super računara, sa konfiguracijom od stotine hiljada procesora. Power 775 u kombinaciji sa specijalizovanim IBM softverom, ovaj sistem je dizajniran da obavlja i predstavlja najnoviju Power tehnologiju.

Idealno radno opterećenje za ovaj sistem su aplikacije koje se bave vremenskom prognozom i modelovanjem klimatskim promena, računska hemija, fizika, računarska tehnika, proračuni dinamike fluida/tečnosti i naftna istraživanja. IBM je je dugo lider u ovim naučnim oblastima, što omogućava inovacije i oblikovanje budućnosti.

Gusto pakovanje 3.072 POWER7® procesorskih jezgri po reku, tako da svako procesorsko jezgro radi na **3.83 GHz**. Fioka 256-core jezgra Power 775 super računara je dizajnirana za brzinu i maksimalne performanse sa mogućnošću grupisanja klastera do 2.048 fioka zajedno, ostvarujući računsku snagu od **524.288** Power7 jezgara koja mogu biti dodeljena, prilikom rešavanja nekih od najvećih svetskih problema. Power 775 može da po rack-u da podrži do **24 TB** ram memorije i **230 TB** prostora za smeštaj podataka, koristeći brze interkonekcije. Snaga Power 775 servera se procenjuje na više od 96 TFLOPS-a po rack-u. Novi radikalni pristup HPC od strane Power 775 sistema predstavlja još jedan korak u evoluciji modularnih klaster rešenja, dizajniranog za rešavanje svetskih problema. Polazeći sa brzinom munje procesora Power7, dodajući ultra gusto pakovanje od **256** jezgara unutar 2U fioke i 100% hlađenje upotrebom nove tehnologije vodenog hlađenja, pruža postizanje vrhunskih performansi čineći ovaj sistem super soničnim automobilom na IT autoputu budućnosti. [28] [29]



Slika 22. IBM Power 775, kućište.

Izvor:

http://www-03.ibm.com/systems/resources/systems_power_hardware_750_692x450.jpg

Tabela 10. Hardverske karakteristike – IBM Power 775

CPU	POWER7 8C 3.84GHz 256 POWER7 cores per drawer
Cores	2048 CEC drawers together makes up 524.288
Linpack Performance (Rmax)	635.111 TFlop/s
Theoretical Peak (Rpeak)	754.188 TFlop/s
Memory	DDR3 128 DIMM slots per drawer Up to 2 TB per drawer, MAX: 24 TB
Network	Up to 17 Quad Port 1 Gb Up to 17 Dual Port 10/100
Integrated Optical Interconnect	Up to 192 50 Gb/s Ports per drawer Up to 128 100 Gb/s Ports per drawer
PCI Expansion / Drawer	16 - 16X PCIe Gen 2, 1 - 8X PCIe Gen 2
Storage	Up to 6 per rack Up to 384 SFF drives per drawer
Cooling	Water (100% Heat capture)
Power	1.386,91 kW
Form factor	2U
Managment	AIX 7.1
RAS features	1. Fully Redundant Power / Cooling / Management 2. Dynamic Processor De-allocation 3. Dynamic Memory chip & lane Sparing 4. Highly Granular Fault Containment 5. Concurrently Maintainable
OS	AIX 7.1, Linux RHEL.11
Price	\$36.485

Izvor: [29] [30]

2.2.3 Uporedni prikaz karakteristika IBM

Tabela 11. Uporedni prikaz karakteristika IBM servera

Server Name	Blue Gene/Q 1 node	IBM Power 775
CPU	1.6 GHz 64-bit POWERPC A2, L1:16 KiB + L1d: 16 KiB, L2: 32 MiB	POWER7 8C 3.84GHz 256 POWER7 cores per drawer
Memory	16 GiB DDR3	DDR3 128 DIMM slots per drawer Up to 2 TB per drawer, MAX: 24 TB
Network	InfiniBand Interconnect 48 Fiber optic	InfiniBand Interconnect Up to 17 Quad Port 1 Gb Up to 17 Dual Port 10/100
Storage	SAN, NAS local storage device DS5020 (16 TB)	SAN or NAS Local, max 230 TB Up to 6 per rack Up to 384 SFF drives per drawer
Boot	Front node + login node	Front node + login node
Cooling	Water (100% Heat capture)	Water (100% Heat capture)
System configurations	Linux Batch commands I/O nodes and JUQUEEN	RAS features: 1. Fully Redundant Power / Cooling / Management 2. Dynamic Processor De-allocation 3. Dynamic Memory chip & lane Sparing 4. Highly Granular Fault Containment 5. Concurrently Maintainable
Power	1 node - 55W, 8 racks - 400 kW, 28 racks – 1.4 MW, 96 racks ~ 4.76 MW	1 server Power 775 - 1.386,91 kW
Form factor	8, 28, 96 racks	2U
OS	Red Hat Enterprise Linux 6	AIX 7.1, Linux RHEL.11
Number of servers	6.144	1
Single server	\$ 90.000	\$ 36.485
Price System	Not available	Not available

Izvor: sopstvena izrada, dobijeni podaci na osnovu istraživanja unutar rada.

2.3 Cray Inc.

Cray Inc. je osnovana od strane **Seymour Cray** 1972. godine pod prvobitnim nazivom **Cray Research**. Cray predstavlja staru kompaniju sa novim imenom, novo ime je promenjeno i registrovano u decembru 1987. pod nazivom Cray inc. Razlog za promenu imena je udruživanje imovine sa Tera Computer Company.

Kompanija Cray Inc. se 40 godina bavi razvojem i izradom rešenja za potrebe najkompleksnijih svetskih nauka, inženjeringa i analitičkih izazova. Sedište kompanije se nalazi u Seattle, Washington, dok proizvodni pogoni su locirani u California, Minnesota, Texas i Wisconsin-u, brojeći preko 1.000 zaposlenih širom sveta. Od svog prvog predstavljanja super računara u 1976 godini, konstantno su učestvovali u razvijanju ideja i kompletnih primenjivih HPC rešenja, koja su se koristila u rešavanju **nerešivih** pitanja i problema.

Kao pioniri u oblasti HPC računarstva sarađuju sa korisnicima koji se ozbiljno bave naukom, istraživanjem i rešavanjem problema. Razlog tome je postojanost specijalizovane tri proizvodne linije koje se sačinjavaju od:

1. Super računara
2. Skladištenja podataka sa sistemom za menadžment
3. rešenja za analizu podataka

U zavisnosti od potrebe klijenta, imaju u ponudi pojedinačan ili integrisani sistem sa kompletnim rešenjem za njihove potrebe. Istraživanje i razvoj kompanije je vođen fleksibilnom vizijom super računara budućnosti fokusirajući se na pružanje inovativnih proizvoda koji integrišu u jedinstvenu arhitekturu, različite tehnologije i naučne oblasti omogućavajući korisnicima prevazilaženje današnjih ograničenja računarske snage, ispunjavajući sve veću potražnju za performansama. [31] Cray Inc. u svoj ponudi super računara poseduje dve serije sistema XC i CS, pri čemu ćemo izvršiti analizu samo XC sistema:

1. XC - adaptive supercomputing architecture
 - I. **Cray XC40.**
 - II. **Cray XC40-AC.**
2. CS - cluster supercomputers
 - I. **Cray CS400-AC.**
 - II. **Cray CS400-LC.**

2.3.1 XC SERIES

XC serija donosi prilagodljivu arhitekturu super računara, koja nudi izuzetnu skalabilnost i održivost performansi. Fleksibilnost XC serije obezbeđuje korisnicima precizno konfigurisanje mašine, koja će zadovoljiti njihove specifične potrebe koje će i u budućnosti moći da nadgrade i poboljšaju.

Cray® XC40™ i XC40-AC™ pokreće Intel® Xeon®, a za komunikaciju se koristi **Aries interconnect i Dragonfly mrežnu topologiju**, pružajući malo kašnjenje i prilagodljivi globalni protok zadovoljavajući najizazovnije petaFLOPS aplikacije.

Prilikom upotrebe ekstremnog skaliranja Cray XC40 koristi se poprečni protok vazduha u kombinaciji sa sistemom za vodeno hlađenje. Kombinovano hlađenje pruža gustinu ormana, bez zahteva za smeštajem dodatne rashladne tečnosti ili dodatnih rack ormana. Centralizovana mrežna topologija smanjuje troškove, čineći sistem kompatibilnim sa različitim računskim tehnologijama. [31] [32]

2.3.1.1 Cray XC40

Cray XC40 kombinuje mogućnosti najnovijih Intel Xeon-Phi 5120D (platforma Haswell) procesora i NVIDIA Tesla grafičkih kartica, stvarajući CPU i GPU klastere, koji se povezuju u jednu celinu (**Aries interconnect platforma** i **Dragonfly** topologija). Cray kao i većina konkurentskih kompanija koristi sistem vodenog hlađenja. **CPU klaster** Cray XC40 (Slika 23.) radi na **2.5 GHz** i sadrži **1376 nodova**, pri čemu svaki ima **2 CPU socket-a** sa **12 jezgara**, **128GB RAM-a** i konekciju sa Aries interconnect platformom. Cray XC40 sadrži dva klastera zasnovana na akceleratorima (eng. Accelerator based clusters):

1. **Prvi akcelerator** je zadužen za povezivanje sa NVIDIA GPU karticama, koje sadrže **44 čvora**.
2. **Drugi akcelerator** sa Intel Xeon-Phi ostvaruje povezivanje **48 čvorova**, i čini klaster akcelerator.

Sistem može da podrži maksimalnih **384** Intel procesora unutar jednog kućišta, sa **256GB** RAM memorije i propusnosti od **137GBps**. Svaki node na klaster akceleratoru poseduje **Intel IvyBridge** na 2.4 GHz, 1xCPU socket-om i 12 jezgara zajedno sa 1xGPU ili 1x Phi card sa 64GB RAM-a. GPU kartica koja se koristi u ovom sistemu je **Nvidia Tesla K40** sa **2880 jezgara** i **12GB** RAM memorije, ili ako je u pitanju **Intel Xeon-Phi 5120D** koji sadrži **60 jezgara** i **8GB** RAM memorije. Cray za svoje sisteme skladištenja podataka koristi **Lustre file sistem** koji obezbeđuje brzinu i propusnost podataka HPC sistema. Razlog za upotrebom Lustre file sistema je u upotrebi prilagođenog Linux-a, pod nazivom **Cray Linux Environment** koji koristi specifične arhitekture: Cray, Intel kompajlera (eng. Compilers) i kompajlera otvorenog koda (eng. Open source) zasnovane na GNU publikaciji. Unutar sistema su integrisane paralelne biblioteke kao što su: **OpenMPI, MPI, CUDA i Intel softverskog klastera**. Proširena podrška integrisana unutar sistema sa širokim spektrom paralelnih naučnih i matematičkih biblioteka kao što su: **BLAS, LAPACK, Scalapack, fftv, hdf5, netcdf, PETSc, Trilinos** itd.

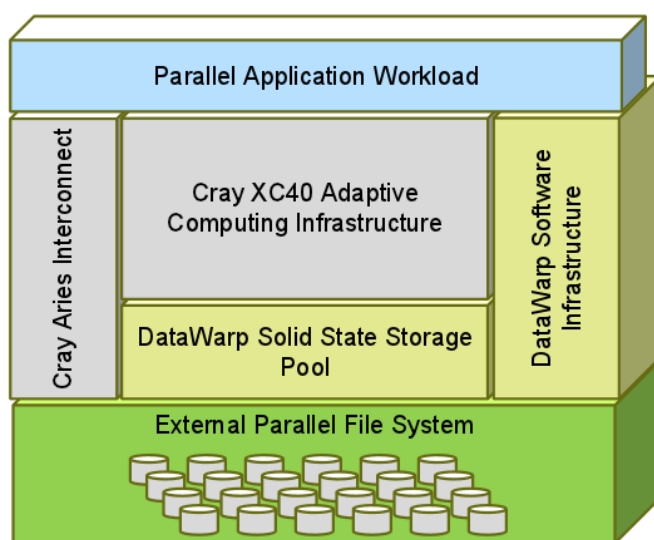


Slika 23. Sastavljen Cray XC40

Izvor: http://regmedia.co.uk/2014/09/30/cray_xc40_900.jpg

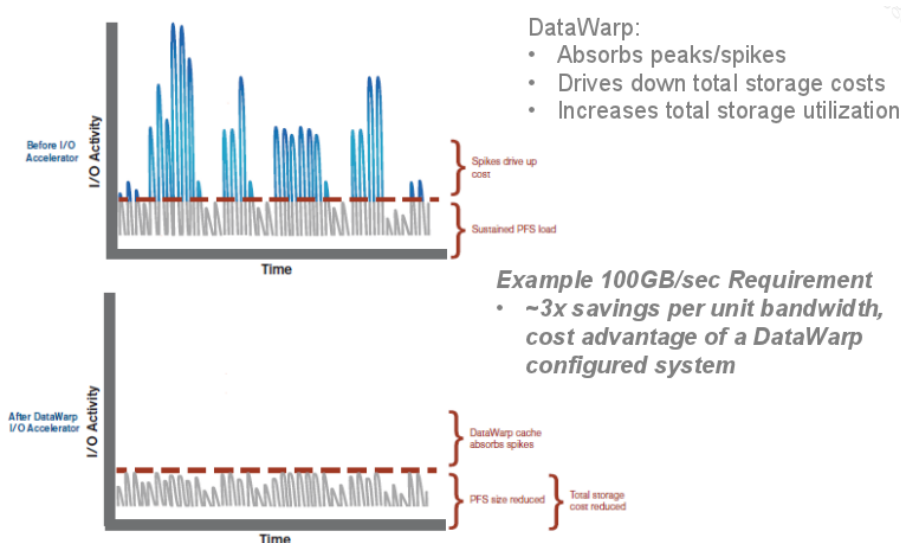
Razlog za integracijom biblioteka jeste da se korisnicima olakša paralelni razvoj programa u najnovijoj verziji DDT softvera za paralelno otklanjanje grešaka i profilisanje koje je omogućeno na sistemu. Pristupanje i upotreba sistema se vrši upotrebom login nodova, upotrebom korisničkog imena i domena na kojem se koristi sistem (ime.prezime@imedomena.com). Unutar login nodova moguće je pokretanje zadataka i batch operacija. Od korisnika se očekuje, da upotrebom ssh konekcije prosledi login nodovima odgovarajuće radne scenarije i skripte koje će se koristiti unutar **PBSPro batch** softvera. Paralelni sistem datoteka je dostupan korisnicima kroz skripte zadatka (eng. Job scripts) i namenjen je za upotrebu prilikom izvršavanja zadataka.

Cray XC40 izdvaja upotreba DataWarp I/O tehnologije (Slika 24, 25.) koja objedinjuje upotrebu DDR4 ram memorije, Cray Aries Interconnect, DataWarp SSD Storage Pool, DataWarp Software Infrastructure, External Parallel File System, Parallel Application Workload i PCI Express 3.0 interfejsa za povezivanje SSD diskova, što znatno smanjuje distancu (izraženu u nano sekundama) između procesora i SSD diskova koji se koriste za keširanje (eng. caching). Upotrebom PCI Express 3.0 interfejsa se nadmašuje standardni SATA interfejs i omogućava se veća propusnost u čitanju i pisanju podataka. Upotreba DataWarp I/O tehnologije i SSD diskova (eng. solid-state drive) za direktno povezivanje unutar računskih nodova, smanjuje distancu putovanja informacije od hard diska do procesora, poboljšavajući performanse i smanjujući potrošnju električne energije. [33] [34] [35]



Slika 24. Cray DataWarp I/O tehnologija.

Izvor: <http://6lli539m39y3hpkelqsm3c2fg.wpengine.netdna-cdn.com/wp-content/uploads/2014/09/datawarparch.png>



Slika 25. Razlika u brzini prilikom upotrebe DataWarp I/O tehnologije.

Izvor: <http://6lli539m39y3hpkelqsm3c2fg.wpengine.netdna-cdn.com/wp-content/uploads/2014/09/datawarp3.png>

Tabela 12. Hardverske karakteristike - Cray XC40

CPU	64-bit Intel® Xeon® processor E5 family; Intel(R) Xeon(R) CPU E5-2680 v3 (30M Cache, 2.50 GHz) up to 384 per cabinet
Memory	64-256 GB per node
Network Interconnect	1 Aries routing and communications ASIC per 4 compute nodes 48 switch ports per Aries chip (500 GB/s switching capacity per chip) Dragonfly interconnect: low latency, high bandwidth topology Infiniband, 40 and 10 Gigabit Ethernet, Fibre Channel (FC) and Ethernet
Storage	Full line of FC, SAS and IB based disk arrays with support for FC and SATA disk drives, Sonexion® data storage system Lustre, Data Virtualization Service (DVS) allows support for NFS, external Lustre and other file systems Max storage 5.4 PB
Boot	Sonexion® data storage system Lustre, Data Virtualization Service (DVS)
System configurations	Cray System Management Workstation (SMW) Single-system view for system administration System software rollback capability
Power	90 kW per compute cabinet, maximum configuration Support for 480 VAC and 400 VAC computer rooms 6 kW per blower cabinet, 20 AMP at 480 VAC or 16 AMP at 400 VAC (three-phase, ground) System ~1512 KW
Cooling	Water cooled with forced transverse air flow: 6,900 cfm intake
Form factor	H 80.25" x W 35.56" x D 62.00" (compute cabinet) H 80.25" x W 18.00" x D 42.00" (blower cabinet) 3,450 lbs. per compute cabinet—liquid cooled, 243 lbs./square foot floor loading 750 lbs. per blower cabinet
Managment	PBS Professional job management system Moab Adaptive Computing Suite job management system SLURM – Simple Linux Unified Resource Manager
OS	Cray Linux Environment (includes SUSE Linux SLES11, HSS and SMW software), Extreme Scalability Mode (ESM) and Cluster Compatibility Mode (CCM)
Number of servers	Number of compute nodes: 3944, Number of service nodes: 44 Special nodes External Access Nodes, Pre- & Postprocessing Nodes, Remote Visualization Nodes
Price System	Not Available

Izvor: [31] [33]

2.3.1.2 Cray XC40-AC

Cray XC40-AC koristi iste komponente kao i Cray XC40 ostvarujući približno iste performanse sa razlikom u broju servera po rack-u i načinu hlađenja koristeći sistem hlađenja vazduhom i vazdušnog strujanja, za razliku od XC40 koji koristi vodeno hlađenje. XC40-AC poseduje tri vrste Dragonfly mrežne topologije, to jest načina povezivanja:

1. Individualno povezivanje sadržano unutar jednog kućišta, direktno povezuje nodove na sistem eliminišući potrebu za dodatnim kablovima.
2. Pasivna mreža povezuje različite kućišta/kabinete upotrebom električne mreže.
3. Aktivna optička veza koja se koristi prilikom povezivanja na veći HPC sistem.

Tabela 13. Hardverske karakteristike - Cray XC40-AC

CPU	64-bit Intel® Xeon® processor E5 family; Intel(R) Xeon(R) CPU E5-2680 v3 (30M Cache, 2.50 GHz) up to 128 per cabinet
Memory	64-256 GB per node
Network Interconnect	1 Aries routing and communications ASIC per 4 compute nodes 48 switch ports per Aries chip (500 GB/s switching capacity per chip) Dragonfly interconnect: low latency, high bandwidth topology Infiniband, 40 and 10 Gigabit Ethernet, Fibre Channel (FC) and Ethernet
Storage	Full line of FC, SAS and IB based disk arrays with support for FC and SATA disk drives, Sonexion® data storage system Lustre, Data Virtualization Service (DVS) allows support for NFS, external Lustre and other file systems Max storage 5.4 PB
Boot	Sonexion® data storage system Lustre, Data Virtualization Service (DVS)
System configurations	Cray System Management Workstation (SMW) Single-system view for system administration System software rollback capability
Power	30 kW per compute cabinet, maximum configuration Support for 480, 400, 230, 208 and 200 VAC computer rooms
Cooling	Air cooled with bottom to top airflow: 3,000 cfm intake
Form factor	H 80.75" x W 27.65" x D 64.00" 1,500 lbs maximum operational 1,800 lbs maximum shipping
Managment	PBS Professional job management system Moab Adaptive Computing Suite job management system SLURM – Simple Linux Unified Resource Manager
OS	Cray Linux Environment (includes SUSE Linux SLES11, HSS and SMW software), Extreme Scalability Mode (ESM) and Cluster Compatibility Mode (CCM)
Number of servers	Number of compute nodes: 3944, Number of service nodes: 44 Special nodes External Access Nodes, Pre- & Postprocessing Nodes, Remote Visualization Nodes
Price System	Not Available

Izvor: [36]

2.3.2 Uporedni prikaz karakteristika Cray XC

Tabela 14. Uporedni prikaz karakteristika Cray XC servera

Server Name	Cray XC40	Cray XC40-AC
CPU	64-bit Intel® Xeon® processor E5 family; Intel(R) Xeon(R) CPU E5-2680 v3 (30M Cache, 2.50 GHz) up to 384 per cabinet	64-bit Intel® Xeon® processor E5 family, up to 128 per cabinet
Memory	64-256 GB per node Memory bandwidth: up to 137 GB/s per node	64-256 GB per node Memory bandwidth: up to 137 GB/s per node
Network	1 Aries routing and communications ASIC per 4 compute nodes 48 switch ports per Aries chip (500 GB/s switching capacity per chip) Dragonfly interconnect: low latency, high bandwidth topology Infiniband, 40 and 10 Gigabit Ethernet, Fibre Channel (FC) and Ethernet	1 Aries routing and communications ASIC per 4 compute nodes 48 switch ports per Aries chip (500 GB/s switching capacity per chip) Dragonfly interconnect: low latency, high bandwidth topology Infiniband, 40 and 10 Gigabit Ethernet, Fibre Channel (FC) and Ethernet
Storage	Full line of FC, SAS and IB based disk arrays with support for FC and SATA disk drives, Sonexion® data storage system Lustre, Data Virtualization Service (DVS) allows support for NFS, external Lustre and other file systems Max storage 5.4 PB	Full line of FC, SAS and IB based disk arrays with support for FC and SATA disk drives, Sonexion® data storage system Lustre, Data Virtualization Service (DVS) allows support for NFS, external Lustre and other file systems Max storage 5.4 PB
Boot	Sonexion® data storage system Lustre, Data Virtualization Service (DVS)	Cray System Management Workstation (SMW) Single-system view for system administration System software rollback capability
Cooling	Water cooled with forced transverse air flow: 6,900 cfm intake	Air cooled with bottom to top airflow: 3,000 cfm intake
System configurations	Cray System Management Workstation (SMW) Single-system view for system administration System software rollback capability	Cray System Management Workstation (SMW) Single-system view for system administration System software rollback capability
Power	90 kW per compute cabinet, maximum configuration Support for 480 VAC and 400 VAC computer rooms 6 kW per blower cabinet, 20 AMP at 480 VAC or 16 AMP at 400 VAC (three-phase, ground) System ~1512 KW	30 kW per compute cabinet, maximum configuration Support for 480, 400, 230, 208 and 200 VAC computer rooms
Form factor	H 80.25" x W 35.56" x D 62.00" (compute cabinet)	H 80.75" x W 27.65" x D 64.00" 1,500 lbs maximum operational

	H 80.25" x W 18.00" x D 42.00" (blower cabinet) 3,450 lbs. per compute cabinet—liquid cooled, 243 lbs./square foot floor loading 750 lbs. per blower cabinet	1,800 lbs maximum shipping
OS	Cray Linux Environment (includes SUSE Linux SLES11, HSS and SMW software) Extreme Scalability Mode (ESM) and Cluster Compatibility Mode (CCM)	Cray Linux® Environment (includes SUSE Linux SLES11, HSS and SMW software) Extreme Scalability Mode (ESM) and Cluster Compatibility Mode (CCM)
Number of servers	Number of compute nodes: 3944, Number of service nodes: 44Special nodes External Access Nodes, Pre- & Postprocessing Nodes, Remote Visualization Nodes	Number of compute nodes: 3944, Number of service nodes: 44Special nodes External Access Nodes, Pre- & Postprocessing Nodes, Remote Visualization Nodes
Price System	Not Available	Not Available

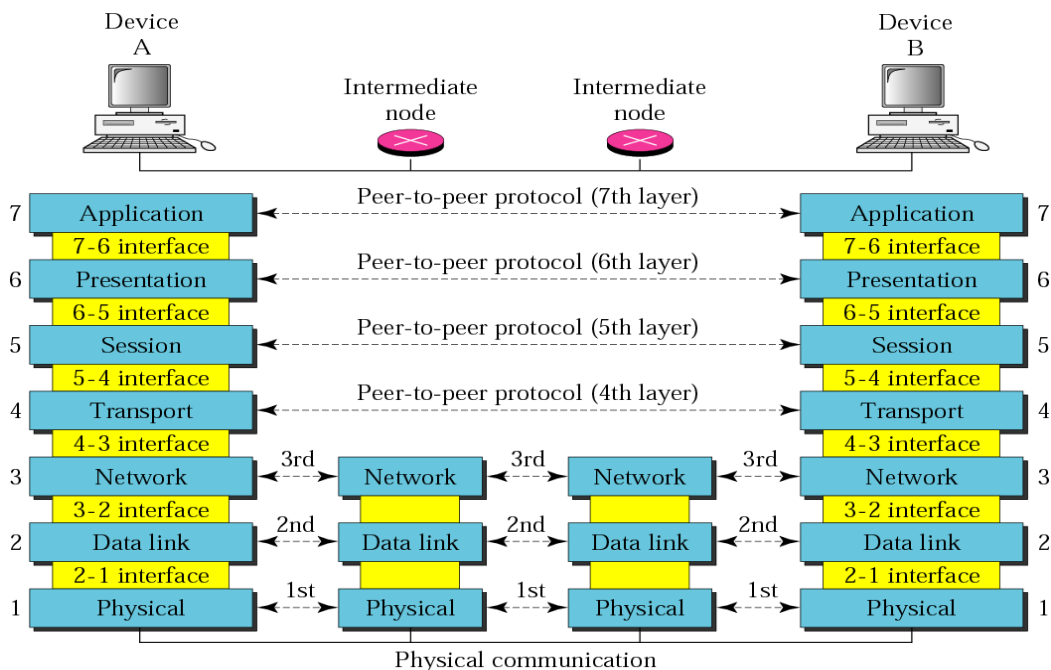
Izvor: sopstvena izrada, dobijeni podaci na osnovu istraživanja unutar rada.

3 Protokoli za komunikaciju HPC sistema

Uspešnost HPC sistema se zasniva na uspešnoj komunikaciji i kvalitetnom hardveru zbog čega je neophodno razumevanje računarskog hardvera, mrežne komunikacije, sigurnosti prenosa podataka, protokola i mogućnosti propusnosti podataka upotrebom određenog protokola. U telekomunikacijama/komunikacijama, protokol i način komuniciranja predstavlja ključan osnov za funkcionisanje sistema i njegovu efikasnost.

Važnost protokola za komunikaciju je način prenosa informacija od tačke **A** do tačke **B** (Slika 1.), ili između više udaljenih tačaka sa njihovim kombinacijama zajedno. Teoretski paketi mogu da putuju kroz beskonačan broj puteva, između tačke **A** i tačke **B**, ali na kraju podaci uvek putuju istim putem od **fizičkog** do **aplikativnog** sloja. Osnov za uspešnu realizaciju komunikacije između dva ili više entiteta, određuje sistem pravila unutar svakog protokola koji se koristi, tako da svaki od protokola određuje način na koji će se vršiti komunikacija: slanje poruka, putanja za slanje, propusnost informacija itd. Na primer TCP/IP protokol da bi ostvario uspešan prenos podataka, primalac mora da primi pakete u tačnom redosledu u kojem su poslali. U slučaju da i jedan paket bude izgubljen iz bilo kog razloga, protokol to može tumačiti kao znak preopterećenosti mreže, smanjujući brzinu prenosa, nakon čega će pokušati da ponovo uspostavi konekciju. Ovo može da bude idealno u nekim situacijama i veoma neefikasno u drugim, jer sam protokol ne poseduje inteligenciju i mogućnost odlučivanja za izvršavanjem određenih promena u komunikaciji. [37]

Mnogi od protokola uključuju upotrebu osnovnih funkcionalnosti TCP/IP protokola, kao što su kontrola paketa (način slanja poruke), detekcija vrste mreže (žičana, bežična), formatiranje poruka za komunikaciju, rukovanje (eng. handshaking), slanje i primanje poruka/paketa, prekidanje sesija ili konekcija, popravka neispravnih i nepotpunih poruka itd. HPC sistem u svojoj osnovi za komunikaciju koristi TCP/IP protokol i njegove varijacije. Na prikazanoj fotografiji (Slika 26.) možemo videti neke od glavnih slojeva TCP/IP protokola prilikom ostvarene komunikacije, i njegovu važnost za ostale protokole koji se oslanjaju na TCP/IP arhitekturu.

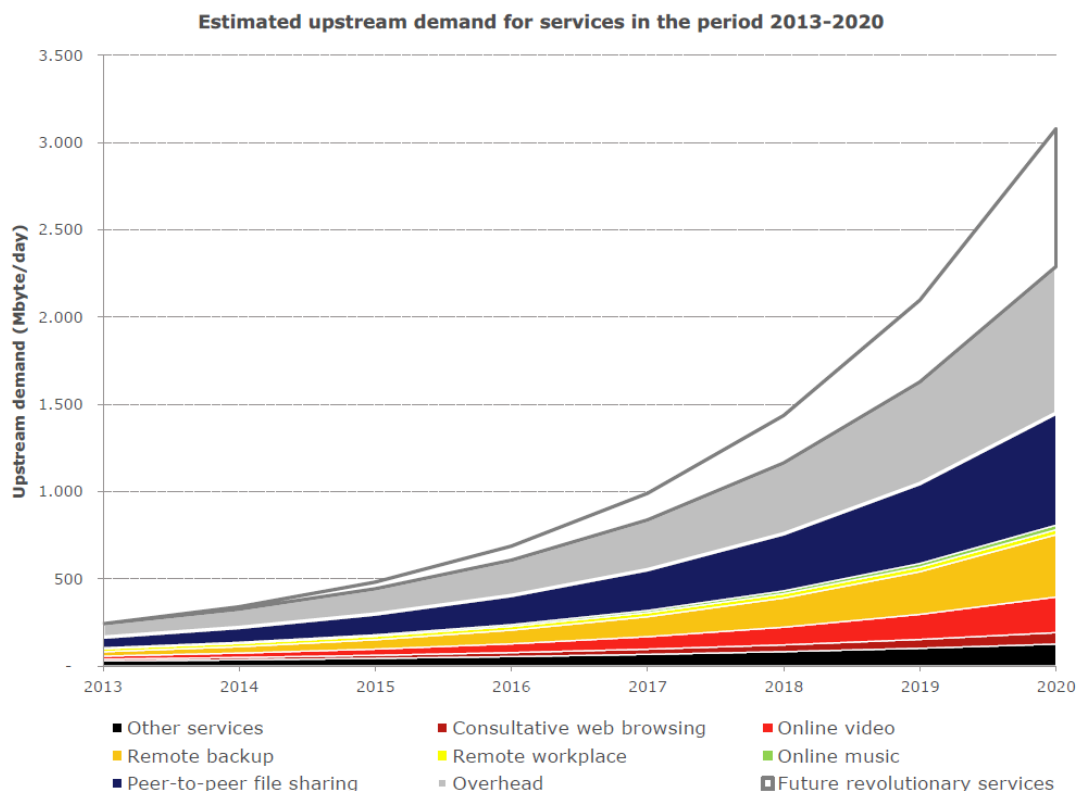


Slika 26. A-B komunikacija OSI model

Izvor: <http://image.slidesharecdn.com/referencemodelsnew-150106001831-conversion-gate01/95/reference-models-in-networks-osi-tcpip-7-638.jpg?cb=1420503604>

Stalna potražnja za što brzim prenosom podataka preko računarske mreže je dovela do promene i poboljšanja standarda za komunikaciju, praveći protokole koji će biti u mogućnosti da podrže današnje i sutrašnje zahteve za protokom velike količine mrežnog saobraćaja. Razumevanja protokola i brzine se ogleda u zahtevima i potrebama tržišta, za postizanjem sve većih brzina prenosa podataka.

Porast brzine prenosa (Slika 27.) izračunate su na osnovu potražnje i obima dnevnog saobraćaja koji se najviše koristi (u vreme u kome je kojim se prenosi). Obim prometa je procenjen po grupama korisnika i po Servisnim kategorijama pokazujući koje usluge će imati najveće zahteve po pitanju mrežnog saobraćaja, pri čemu razmena video materijala se svrstava među prve tri kategorije. [38] [39] [40]



Slika 27. Projektovana prosečna količina saobraćaja za period od 2013-2020.

Izvor: [40]

Razvoj i potražnja za većim upload brzinama je povezana sa brzinom preuzimanja podataka. Između ostalog download brzine su kritičnije za krajnje korisnike koji često "čekaju" da bi videli sadržaj ili koristili neku mrežnu aplikaciju. Međutim, mnogo je kritičnija upload brzina za direktnu distribuciju podataka. Potražnja za propusnim opsegom se očekuje da će eksponencijalno rasti tokom narednih godina (Slika 27.).

Protokoli koje se koriste u komunikaciji, raspoređeni u današnjim najzahtevnijim mrežama moraju da poseduju mogućnost propusnosti velike količine podataka, pri čemu moraju da poseduju standardizovane specifikacije za postizanje sve veće brzine prenosa, tako da podržavaju i prethodne verzije protokola koje koriste u svojoj implementaciji. [39] [40] Fizička infrastruktura koja podržava ove prenose podataka kao što su Interconnect, InfiniBand, Torus 3D-5D, optičke mreže itd. predstavlja ključnu ulogu u povezivanju (fizičkog) sistema, obezbeđujući potrebnu brzinu i propusnost podataka sa mogućnošću unapređivanja u budućnosti. Unutar HPC sistema posvoji veliki broj protokola i tehnologija koji se koristi za komunikaciju i ostvarivanje performansi tako da ćemo obraditi najznačajnije protokole koji su sadržani u skoro svakom HPC sistemu.

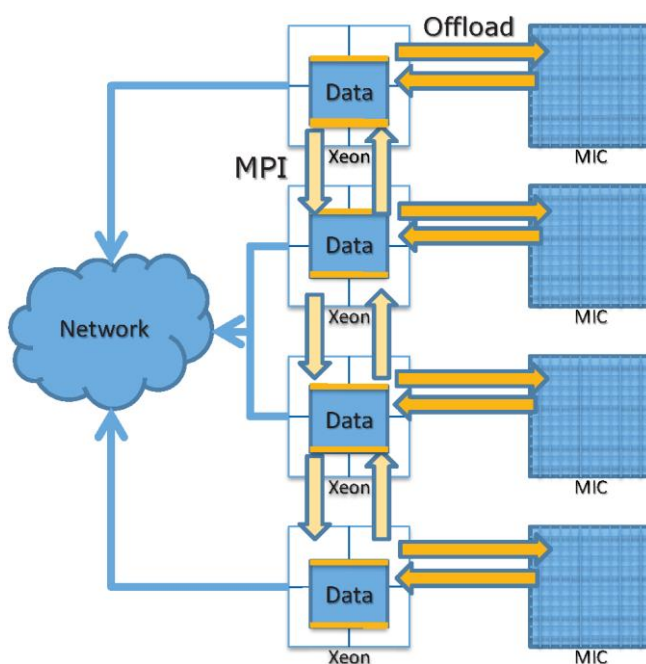
3.1 MPI

MPI kao osnovni standard za prosleđivanje poruka (eng. Message Passing Interface) je standardizovana biblioteka za prosledjivanje poruka, zasnovana na konsenzusu MPI foruma kojeg čini preko 40 organizacija uključujući proizvođače, istraživače, programere softverskih biblioteka i krajnje korisnike sistema za komunikaciju. Cilj ovog standarda za komunikaciju jeste u uspostavljanju standardizovanog portabilnog, efikasnog i fleksibilnog pristupa za prosleđivanje poruka. MPI je široko rasprostranjen za potrebe pisanja programa za HPC sisteme. Kao takav MPI je prvi standardizovani protokol za komunikaciju, koji ne zavisi od proizvođača softvera, operativnog sistema i računarske arhitekture x86 ili x64. Kao standard za komunikaciju na HPC sistemima, nije sastavni deo IEEE ili ISO standarda, ali je postao „industrijski standard“ za pisanje i prosleđivanje poruka između programa na HPC platformama. Prednosti upotrebe MPI su u razvoju poruka, koje odgovaraju dizajniranim ciljevima portabilnosti, efikasnosti i fleksibilnosti.

Primer komunikacije HPC sistema upotrebom MPI protokola je prikazan na (Slika 28.), na kojoj možemo videti razmenu poruka između 4 servera upotrebom računarske mreže, koji su povezani u klaster. Razlog za upotrebu MPI protokola je u mogućnosti direktnog prosleđivanja poruka na više od 1.000 procesora. [41] [42]

3.1.1 OpenMPI

OpenMPI je nastao kao posledica spajanja osnovnih verzija (MPI protokola) kao što su: LAM/MPI, LAM, La-MPI i FT-MPI. Institucije koje učestvuju u razvoju, programiranju i izgradnji OpenMPI protokola, poseduju značajno iskustvo u implementaciji protokola unutar više različitih operativnih sistema fokusirajući se na stvaranje brze i sigurnije komunikacije, prilikom upotrebe HPC softvera koji podržava MPI 1.2 i 2.0 specifikacije i rad na više procesorskih niti (eng. multi-thread) koji upotrebom mrežnih ili fizičkih interconnect-cija stvaraju niz paralelnih operacija, koje dele procesor, memoriju i sve ostale resurse u cilju ostvarivanja što većeg broja TFLOPS-a. OpenMPI obezbeđuje osnovne usluge za početak i upravljanje paralelnim aplikacijama bez obzira na okruženje, koje može biti interaktivno ili neinteraktivno, stvarajući mogućnost portabilnosti i komunikacije na više različitih sistemima (Linux, Windows). Projekat OpenMPI je protokol/interfejs otvorenog koda koji se razvija i održava od strane konzorcijuma akademika, istraživača i industrijskih partnera tako da imaju mogućnost kombinovanja stručnosti eksperata, tehnologija i resursa, koji su dostupni unutar zajednice sistema visokih performansi u cilju izgradnje najbolje MPI biblioteke. Sam protokol poseduje određene prednosti za sistem i proizvođače softvera, aplikacija i istraživače računarskih nauka. [43] [41]



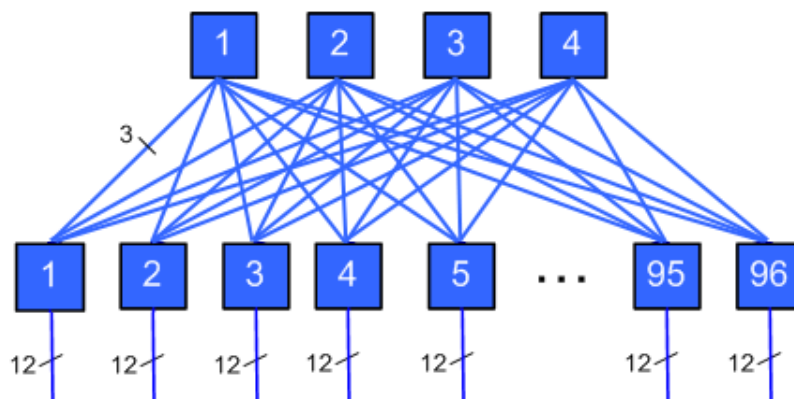
Slika 28. Paralelna komunikacija upotrebom MPI protokola

Izvor: [41]

3.2 Interconnect

Interconnect predstavlja mogućnost globalne dostupnosti informacija koje se nalaze na jednom ili više računara upotrebom računarske mreže ili direktne veze (Slika 29.). Sama komunikacija ovakvog tipa predstavlja veliko opterećenje za bilo koju mrežu. Razlog za to je propusnost opsega medijuma/kabla i protokola koji se koristite za komunikaciju. Vreme pristupanja memoriji na udaljenom računaru može biti dosta različito, što značajno može degradirati mrežnu komunikaciju stvarajući veliko kašnjenje i prekidanje uspostavljenih konekcija. Posledica velikog kašnjenja i nemogućnost prosleđivanja informacija HPC sistemu upotrebom MPI protokola, narušava doslednost, brzinu dobijanja informacija.

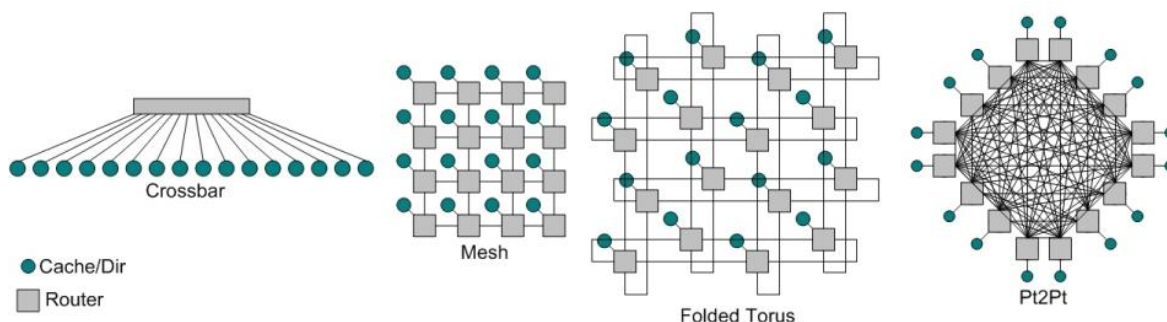
Interconnect uređaji služe za rešavanje problema, međusobne povezanosti različitih mreža i računara koji se nalaze u različitim LAN, WAN mrežama. Fizičke mreže mogu biti povezane na različite načine, upotrebom različitih uređaja [37]. Najčešće metode za povezivanje su upotrebom Repetitera, Mostova, Ruter, komutatora (eng. Switch), konvertera protokola ili direktnim povezivanjem između servera, interconnect tehnologijom upotrebom fizičke konekcije na matičnoj ploči.



Slika 29. Primer Interconnect komunikacije

Izvor: https://computing.llnl.gov/tutorials/linux_clusters/images/1152wayIB.gif

Jedan od glavnih izazova sa kojima se susreću dizajneri digitalnih sistema, jeste optimizacija komunikacije i međusobne povezanosti između ovih sistemskih komponenti (Slika 30.). Interconnect mreže pružaju atraktivno i ekonomično rešenje kada je u pitanju komunikacija i postizanje zavidnih performansi HPC sistema. Trenutni trendovi povećanja brzine govore da će u budućnosti propusnost informacija biti još problematičnija prilikom dizajniranja budućih generacija računara. Shodno tome, jedan interconnect mrežni ruter će u narednim godinama biti veoma poželjan, zbog svoje brzine i direktnog povezivanja na računarsku mrežu. [44]



Slika 30. Primer povezanosti rutera, Interconnect mreža

Izvor: <https://i1.wp.com/commons/9/96/NetworkTopologies.png>

Interconnect mreže pružaju atraktivno rešenje za direktnu komunikaciju između uređaja, i postaju sve prisutnije kao integrisano komunikaciono rešenje. Koriste se u vidu sistema ugrađenih na čipu (eng. SoC On the Chip) matične ploče ili posebne PCI kartice sa glavnom ulogom obrade podataka između procesora, memorije, ulaza/izlaza podataka, rutiranja saobraćaja, u cilju zamene povezivanja fizičkim kablovima (fizičkom žicom).

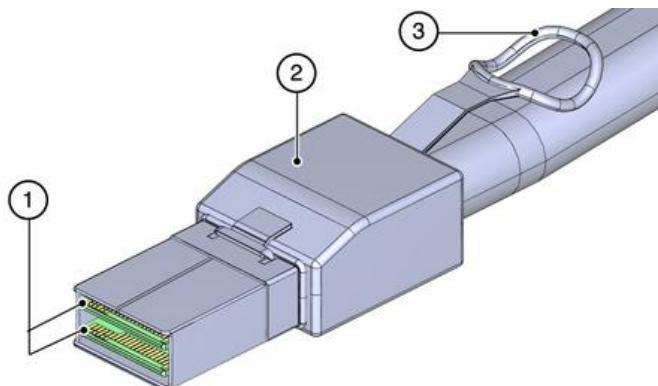
Složenost i integracija sistema nastavlja da raste usmeravajući dizajnere na pravljenje algoritama za usmeravanje paketa podataka upotrebom interconnect arhitekture. Dobar dizajn za povezivanje mrežnih uređaja, koji će omogućiti efikasno korišćenje ograničenih komunikacionih resursa između samih klijenata i servera, čineći sistem lakšim za projektovanje, optimizaciju i ispravljanje grešaka. [44]

Osnovni principi interconnect mreže su relativno jednostavni, i lako je osmisлити efikasnu mrežu koja ispunjava sve zahteve za datu primenu. Međutim, ako se osnovni principi nisu shvatili sistem će loše raditi, usled lošeg izbora mrežne topologije ili algoritma za rutiranje, stvarajući usko grlo u kontroli protoka podataka i nemogućnosti ostvarivanja maksimalnih performansi sistema zbog čega kao dizajneri našeg budućeg sistema, moramo da budemo upoznati sa brojem računara, servera i svih ostalih uređaja koji učestvuju u komunikaciji HPC sistema.

3.3 InfiniBand

InfiniBand kao nova standardizovana arhitektura za komunikaciju, specijalno dizajnirana da podrži I/O vezu visoke propusnosti sa celokupnom Internet/Intranet infrastrukturom. Podržan je od strane svih glavnih OEM proizvođača servera, kao sredstvo za proširivanje izvan standardnih komunikacionih granica unutar jednog servera, stvarajući novu generaciju I/O standarda povezanosti. Po prvi put visoki volumen, industrijski standard I/O povezanosti, proširuje ulogu tradicionalnog sistema u kutiji (eng. in the box). InfiniBand je jedinstven u pružanju oba sistema iz kutije i drugog rešenja, spoljne interconnect operacije ostvarujući protok iz "kutije" tako da obezbeđuje vezu na način koji je ranije bio rezervisan samo za tradicionalne mrežne interkonekcije. Ova unifikacija I/O i sistemskog dela mreže, zahteva novu arhitekturu koja podržava potrebe prethodno odvojenih domena. Konkurencija između Ethernet i InfiniBand tehnologija se odvija u više oblasti savremenih informacionih

tehnologija sa najvećim udelom u HPC sistemima i okruženjima za laboratorijsko testiranje, analiziranje i usavršavanje tehnologija. Obe tehnologije mogu da rade na optičkim i bakarnim žicama upotrebom odgovarajućeg standardizovanog konektora za povezivanje (Slika 31,32.). Primer ostvarene komunikacije je prikazan na Slika 34. [45] [46]

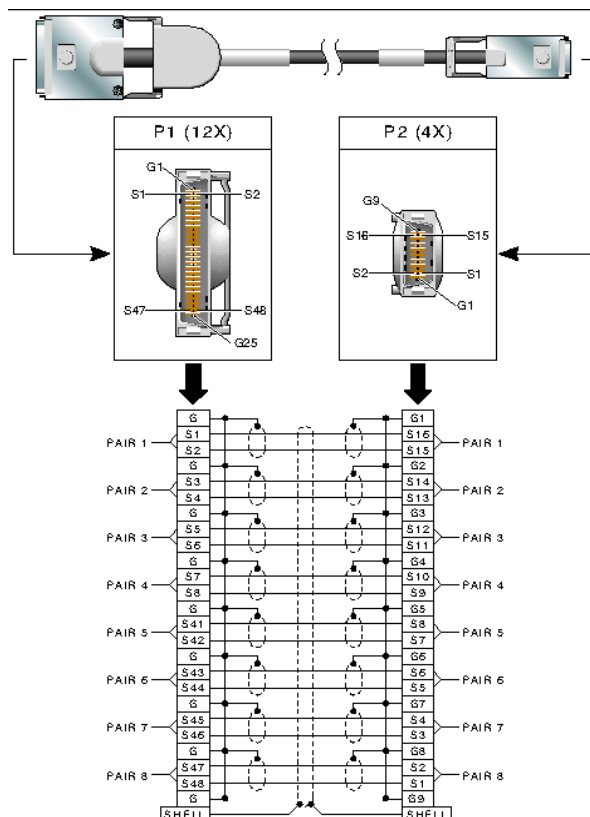


Slika 31. InfiniBand konektor sa kablom.

Izvor: http://docs.oracle.com/cd/E19632-01/835-0783-03/figures/126869IB_cabl_features.jpg

Legenda oznaka (Slika 6.):

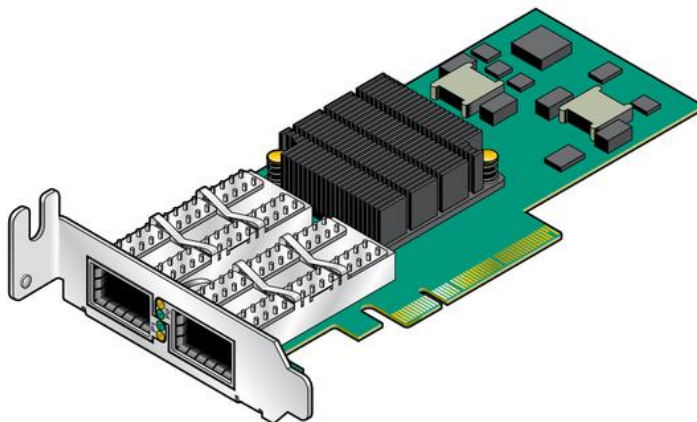
1. Konekcione pločice.
2. Štit.
3. Kaiš za izvlačenje kablova.



Slika 32. InfiniBand kabl i način spajanja pinova.

Izvor: <http://docs.oracle.com/cd/E19049-01/pcix.adapt/819-1280-10/figures/appA-3.gif>

InfiniBand kartica (Slika 33.) se spaja na standardni PCI-Express slot računara koji poseduje treću verziju PCI Express slota. Sama kartica može da poseduje više konekcionih portova, koji upotrebom 12X EDR (Tabela 15.) mogu da ostvare **trenutni maksimalni protok od 384 Gbps**.



Slika 33. InfiniBand PCI-E kartica

Izvor: http://docs.oracle.com/cd/E40985_01/html/E40987/figures/126604.jpg

Tabela 15. Brzine prenosa InfiniBand arhitekture.

Name	Speed ^a	Data rate	Fully duplexed rate
1X	2.5 Gbps	2 Gbps	4 Gbps
4X	10 Gbps	8 Gbps	16 Gbps
8X	20 Gbps	16 Gbps	32 Gbps
12X	30 Gbps	24 Gbps	48 Gbps
1X DDR	5 Gbps	4 Gbps	8 Gbps
4X DDR	20 Gbps	16 Gbps	32 Gbps
8X DDR	40 Gbps	32 Gbps	64 Gbps
12X DDR	60 Gbps	48 Gbps	96 Gbps
1X QDR	10 Gbps	8 Gbps	16 Gbps
4X QDR	40 Gbps	32 Gbps	64 Gbps
8X QDR	80 Gbps	64 Gbps	128 Gbps
12X QDR	120 Gbps	96 Gbps	192 Gbps
1X EDR	20 Gbps	16 Gbps	32 Gbps
4X EDR	80 Gbps	64 Gbps	128 Gbps
8X EDR	160 Gbps	128 Gbps	256 Gbps
12X EDR	240 Gbps	192 Gbps	384 Gbps

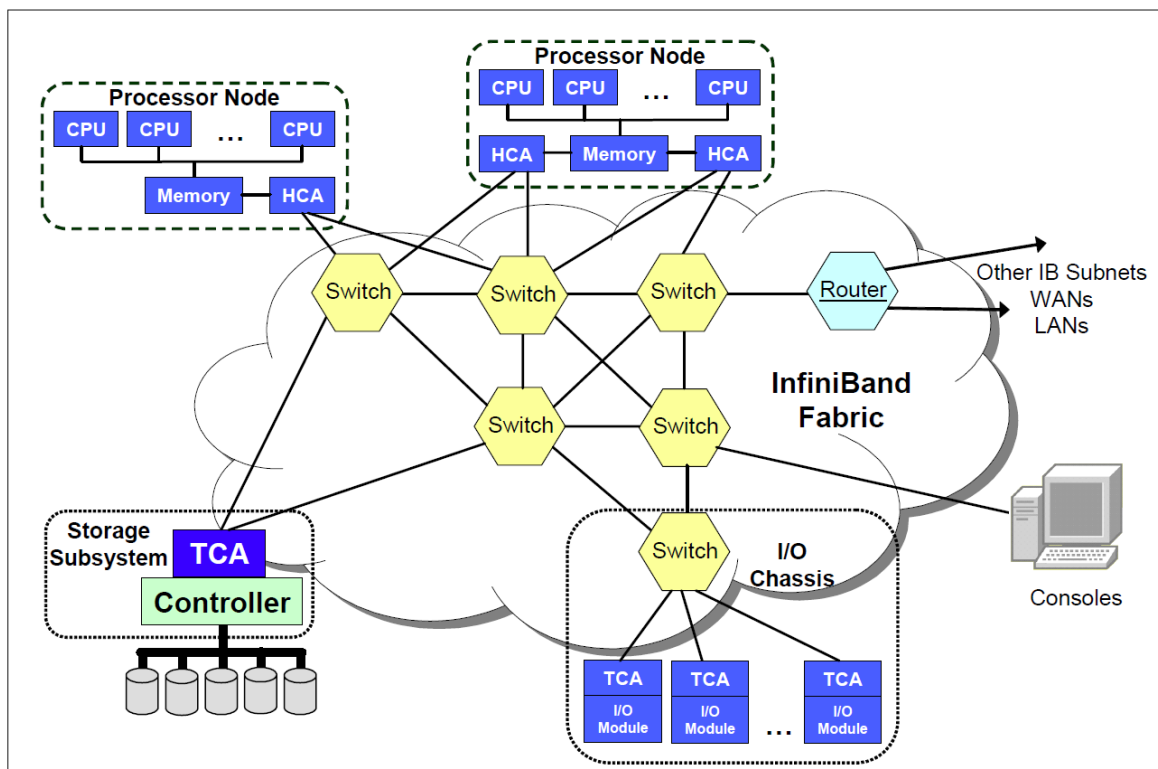
Izvor: [46]

Ono što razlikuje Ethernet i InfiniBand jeste postojanost InfiniBand Trade Association (IBTA) organizacije, osnovane 1999. godine sa zadatkom promocije InfiniBand specifikacija kroz marketing i odnose sa javnošću, sprovodeći interoperabilne testove i mnoge druge testove komercijalnih proizvoda (HP, IBM, Cray Inc itd.) što samim tim dovodi do sve većih integracija u više različitih proizvoda podeljenih u kategorijama kablova i uređaja.

IBTA takođe razvija i redovno ažurira drajvere i podršku InfiniBand tehnologijama unutar različitih operativnih sistema. Prema informacijama grupe pokazuje se projekcija povećane potražnje za InfiniBand tehnologijama, što znači da će se propusni opseg u naredne tri godine približiti cifri od 1.000 Gbits / u sekundi. [45] [46]

InfiniBand podržava nekoliko zajedničkih mrežnih topologija koje koriste različite načine za povezivanje. InfiniBand podržava sledeću listu mrežnih topologija:

1. **Fat tree:** Višestruki koren stabla (eng. multi-root tree). Trenutno najpopularnija i najkorišćenija topologija.
2. **2D mesh:** Svaki nod je povezan sa 4 druga noda koji se nalaze na pozitivnoj i negativnoj X i Y osi.
3. **3D mesh:** Svaki nod je povezan sa 6 drugih noda koji se nalaze na pozitivnoj i negativnoj X, Y, Z osi.
4. **2D/3D torus:** Gde se završavaju X, Y i Z ose na 2D/3D torusu/mesh-u vrši se "umotavanje" (eng. wrapped around) i povezivanje na prvi nod.



Slika 34. Komunikacija unutar InfiniBand mreže.

Izvor: [46]

4 Računarska mreža

4.1 HPC topologija

Topologija unutar HPC sistema je osnov logičke povezanosti fizičkih uređaja na mreži. Svaka mrežna topologija određuje način, oblik povezivanja komponenti (nodova, linkova, eksternih uređaja, kablova itd.) koji učestvuju u mrežnoj komunikaciji predstavljajući strukturu mreže, ne samo fizički već i logički. Način na koji će nodovi komunicirati u HPC sistemu se određuje na osnovu topologije mreže pri čemu se unutar fizičke topologije predstavlja fizički raspored (gore pomenutih uređaja), dok se na logičkoj topologiji predstavlja tok informacija između različitih komponenti u sistemu.

Prilikom izbora mrežne topologije moraju se razmotriti određeni faktori sa kojima raspolažemo:

1. Veličina projekta (broj uređaja/komponenti koji se povezuju unutar sistema).
2. Opterećenje mreže (količina očekivanog saobraćaja na mreži, izražena u GB, TB itd.).
3. Budžet koji imamo na raspolaganju za kupovinu mrežne opreme.
4. Potrebna brzina odgovora na mreži (eng. latency)

4.1.1 Fat tree topologija

Fat tree kao najrasprostranjenija topologija HPC sistema pruža najbolje mrežne performanse kada je u pitanju konfiguracija i rad neblokirajuće mreže velikog broja korisnika. Unutar ove topologije, moguće je podešavanje blokiranja određenog saobraćaja i neisporučenih paketa. Fat tree koristi isti propusni opseg za sve linkove, i u većini slučajeva za komunikaciju između switch-eva se koristi isti broj portova/konekcija.

Prilikom dizajniranja i izgradnju ove topologije moraju se poštovati određena pravila:

1. Ne blokirajuća mreža klastera mora biti uravnotežena, što znači da mora imati isti broj linkova na svakom Layer 2 (L2) switch-u i na svakom nivou povezanosti između switch-eva L1, L2, L3. Broj korisnika i mogućnosti propusnosti zavisi od primene HPC-a sistema i opterećenosti mreže.
2. Ako je L2 switch vrši upravljanje komunikacijom, svi L1 switch linkovi koji idu kao L2 switch-u moraju biti ravnomerno raspoređeni. Na primer, ako postoji 6 veza između L1 i L2 switch-a, može se izvršiti distribucija konekcija: 1:1:1:1:1:1, 2:2:2, 3:3 ili svih 6 konekcija zajedno unutar jednog switch-a. Pri čemu se ne smeju ostvariti ne balansirane veze 4:2, 5:1.
3. Mora se paziti na rutiranje, tako da se ne stvaraju putanje koje moraju da putuju gore, dole (eng. traverse up, traverse down) i onda opet gore putem stabla jer se na ovaj način stvaraju petlje pod nazivom “kreditne petlje” (eng. credit loops) koje se unutar klastera mogu manifestovati kao mrtve zatvorene tačke, stvarajući zagušenje i zastoje celokupnog saobraćaja na mreži. Po pravilu ne postoji način, da se izbegnu ove vrste neprekidnih petlji. Ali kao rešenje za ovu vrstu petlji jeste dobro dizajniranje, povezivanje između switch-eva i direktnih konekcija na više switch-eva, tako da se izbegavaju fizičke petlje i nepotrebna rutiranja saobraćaja.
4. Poželjno je koristiti Layer 1 36 portni switch kao glavni switch, koji će upravljati mrežnom komunikacijom unutar L2 switch-eva.

U slučaju da postoji problem prilikom održavanja klastera na stranici proizvođača (<http://www.mellanox.com/clusterconfig/>) moguće je izvršiti analizu konfiguracije (Slika 35.) i dobijanje nekoliko mogućih rešenja koja su primenljiva u realnom okruženju. Stranica proizvođača takođe prikazuje broj potrebnih komponenti kablova, switcheva i maksimalni broj nodova koji mogu da učestvuju u komunikaciji (Slika 36.). Upotrebom ove topologije moguće je povezati maksimalni broj od **20.736 nodova**.

Slika 35. Mellanox InfiniBand Configurator

Izvor: [47]

8-1 Blocking Cross Bisectional Bandwidth (CBB) Configuration

CLUSTER CONFIGURATION TO USE	SWITCH	FULLY CONFIGURED	PARTIALLY CONFIGURED		
			SWITCH COUNT	PORT COUNT PER SWITCH	LINE CARDS PER SWITCH
LEVEL 1 SWITCHES	36-port Switch	648			
LEVEL 2 SWITCHES*	648-port Switch	4			

* In some scenarios, the cluster could have Level 2 switches in which the number of ports required are different between the various switches. This optimization is accounted for here. If your configuration output has two groups of switches specified, then your cluster can have two groups of switch types, in which one of the groups has less port count than the other group.

CABLES REQUIRED	
Node to Level 1 Switch Cables	20736
Level 1 Switch to Level 2 Switch Cables	2592

Please fill out the form below to receive complete cluster configuration and OEM-specific product part numbers (SKU's).

CONTACT INFORMATION

FIRST NAME:

LAST NAME:

COMPANY:

PHONE NUMBER:

EMAIL ADDRESS:

END USER / OPPORTUNITY:

[EMAIL ME](#)

© 2015 Mellanox Technologies. All Rights Reserved.

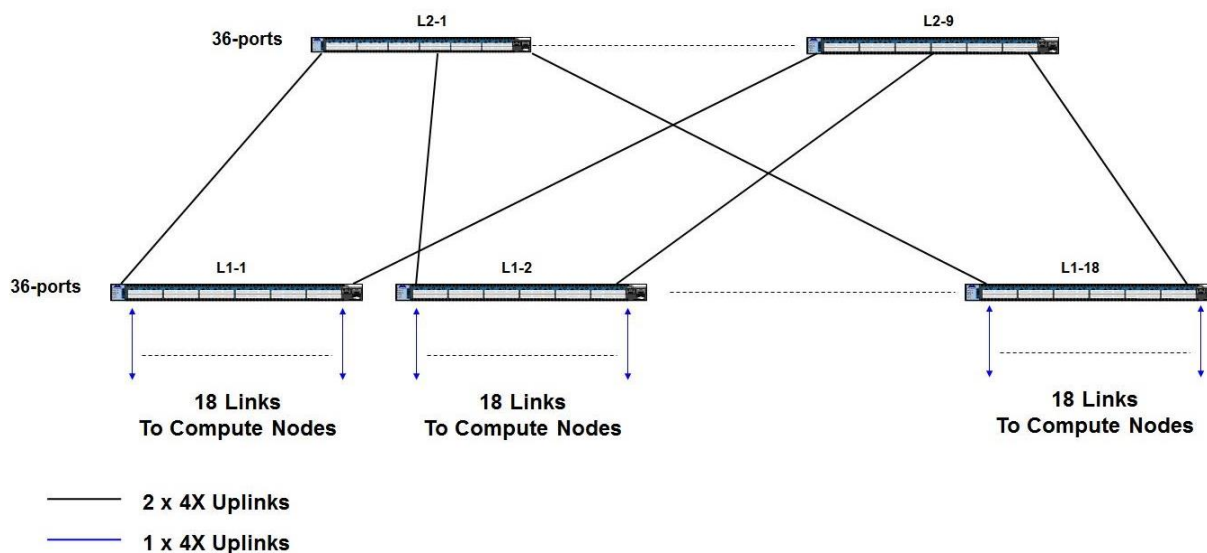
Slika 36. Mellanox InfiniBand Configurator, Bandwidth (CBB) Configuration

Izvor: [47]

Veličina klastera može zahtevati uslove, koji prelaze maksimalni kapacitet neblokirajućih portova Fat Tree topologije (Slika 37.) tako da u zavisnosti od broja portova na switch-u, imamo sledeće primere problema:

1. Ako klaster zahteva upotrebu samo **36 portova**, može se izvršiti osnovna realizacija topologije. Čim ulov prelazi 36 portova, jedan switch mora da kreira **2 nivoa** Fat Tree stabla.
2. U slučaju kada je potrebno **48 porta**, realizacija se vrši sa dva 36 portna switch-a sa odnosom 1:2 što podrazumeva postojanje i kretanje komunikacionih parova od “odredišta do izvora i nazad do odredišta” za čiju komunikaciju je zadužen jedan switch, koji prenosi celokupno komunikaciju upotrebom 2 komunikaciona čvora/linka. Klaster koji koristi 48 portova se može realizovati sa 2 switch-a bez potrebe za upotrebom većeg broja switch-eva.
3. U slučaju kada je potrebno **72 porta**, može se postići potpuna ne blokirajuća Fat Tree topologija, ali jedan switch mora imati **6 portova** koji biti korišćeni za dalje povezivanje.
4. U slučaju kada je potrebno **648 porta** pravi se puna ne blokirajuća mreža na **2 nivoa** koja uključuje prethodne mrežne postavke od **36, 48, 72**.

Kada se pređe granica od 648 portova na 2 nivoa pune neblokirajuće mreže, klaster zahteva veoma pažljivo planiranje za dalje povezivanje na postojeću arhitekturu. Isti koncept se može primeniti i na klastere koji koriste više od gore navedenog broja pri čemu se mora imati na umu validnost i povezanost konfiguracije, kada je u pitanju topologija “prstena” koja podrazumeva 3 switch-a, koji mogu stvoriti beskonačne petlje i zastoj celokupne mreže.



Slika 37. Fat-Tree topologija, 324 noda i 36-portni switch-evi.

Izvor: [47]

4.2 Klaster

Klaster (eng. Cluster) kao jedna složena celina i kostur celokupnog sistema visokih performansi sa svojom mrežom povezanih računara, omogućava radnim stanicama (eng. Nodes) razmenjivanje podataka sa velikim brojem servera (Slika 38.), sve dok su povezani sa glavnim računarom (eng. Head Node) koji predstavlja sve povezane računarske nodove. Mreže su veoma važne za klastere zbog podataka koji prenose, posebno imajući u vidu vrednosti posedovanja prave računarske mreže koja može da primi i prenese celokupan saobraćaj. Od vitalnog značaja je dobro funkcionisanje i lako upravljanje klasterom, zbog čega većina HPC sistema poseduje minimalno 2 mrežne kartice/porta za komunikaciju. Kako bi HPC sistem mogao da postigne maksimalne performanse u svojoj osnovi mora da sadrži:

1. Kontrolu saobraćaja prilikom smeštaja podataka (eng. Storage traffic).
2. Menadžment saobraćaja (eng. Management traffic).
3. Menadžment fotografija ili paketa koji su prosleđeni računskim nodovima, od strane master noda.
4. IPMI saobraćaj mora da bude TCP/IP (eng. IPMI traffic).
5. Nadzor saobraćaja između klijenata i master noda.
6. Izveštavanje o količini saobraćaja i broju konekcija od klijenata ka master nodu.
7. SSH konekcija koja se ostvaruje upotrebom TCP/IP protokola.
8. Zakazivanje protoka dozvoljenog mrežnog saobraćaja, većinom upotrebom TCP/IP protokola.

Formula za izračunavanje performansi računskih nodova i celokupnog klastera, izraženog u FLOPS-ima se računa na sledeći način:

$$\text{Node performance in FLOPS} = (\text{CPU speed in Hz}) \times (\text{number of CPU cores}) \times (\text{CPU instruction per cycle}) \times (\text{number of CPUs per node})$$

Formula 1. Formula za izračunavanje brzine izražene u FLOPS-ima.

Izvor: [47]

Na primer, Intel Dual-CPU server zasnovan na procesoru Intel E5-2600 (2.9GHz 8-cores) po formuli može da ostvari sledeće performanse:

$$\text{Node performance in FLOPS} = \text{CPUs: } 2.9 \times 8 \times 8 \times 2 = 371.2 \text{ GFLOPS po serveru}$$

Formula 2. Performanse po serveru za procesor Intel E5-2600.

Izvor: [47]

Broj instrukcija po ciklusu za procesor Intel Dual-CPU E5-2600 je jednak broju 8. Da bi se izračunale maksimalne performanse klastera, vrši se množenje sa brojem nodova, kako bi se dobila približna teoretska brzina. Na primerima 1,2 i Slika 13. možemo videti maksimalne teoretske performanse u zavisnosti od broja nodova i mrežne konekcije.

1. **72-node** fat-tree (upotrebom 6 switch-eva) klaster može da ostvari:

$$371.2 \text{ GFLOPS} \times 72 \text{ (nodes)} = 26,726 \text{ GFLOPS} = \sim \mathbf{27 \text{ TFLOPS}}$$

2. **648-node** fat-tree (upotrebom 54 switch-a) klaster može da ostvari:

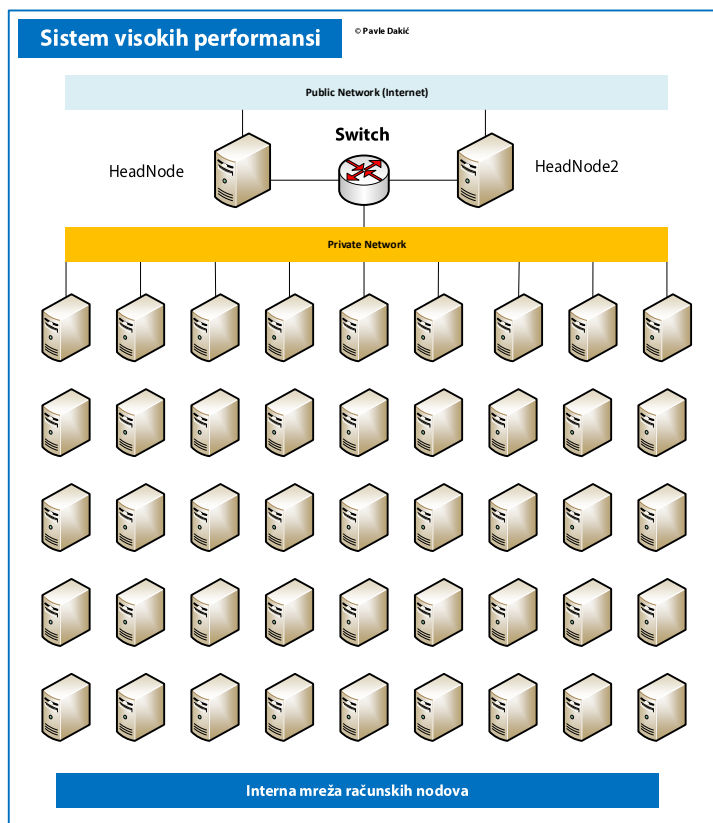
$$371.2 \text{ GFLOPS} \times 648 \text{ (nodes)} = 240,537 \text{ GFLOPS} = \sim \mathbf{241 \text{ TFLOPS}}$$

Cluster Size	Theoretical Performance (100%)	1GbE Network (50%)	10GbE Network (70%)	FDR InfiniBand Network (90%)	Units
72-Node cluster	27	13.5	19	24.3	TFLOPS
324-Node cluster	120	60	84	108	TFLOPS
648-Node cluster	241	120.5	169	217	TFLOPS
1296 Node cluster	481	240	337	433	TFLOPS
1944 Node cluster	722	361	505	650	TFLOPS
3888 Node cluster	1444	722	1011	1300	TFLOPS

Slika 38. Teoretske performanse HPC klastera

Izvor: [47]

U slučaju da klaster sadrži više od **648 nodova**, onda mora sadržati **najmanje 3 nivoa hijerarhije**. Stvarne performanse klastera koje će se ostvariti, najviše zavise od Interconnect operacija između računskih nodova u proseku za 50% se smanjuju upotrebom 1 GbE (eng. Gigabit Ethernet) konekcije. Dok upotrebom 10GbE konekcija, performanse se smanjuju za 30% pri čemu InfiniBand ostvaruje gubitak od samo 10%, što znači 90% efikasnosti sistema naspram drugih mrežnih konekcija (Slika 38.). [47]



Slika 39. HPC klaster

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

4.3 Balansiranje saobraćaja

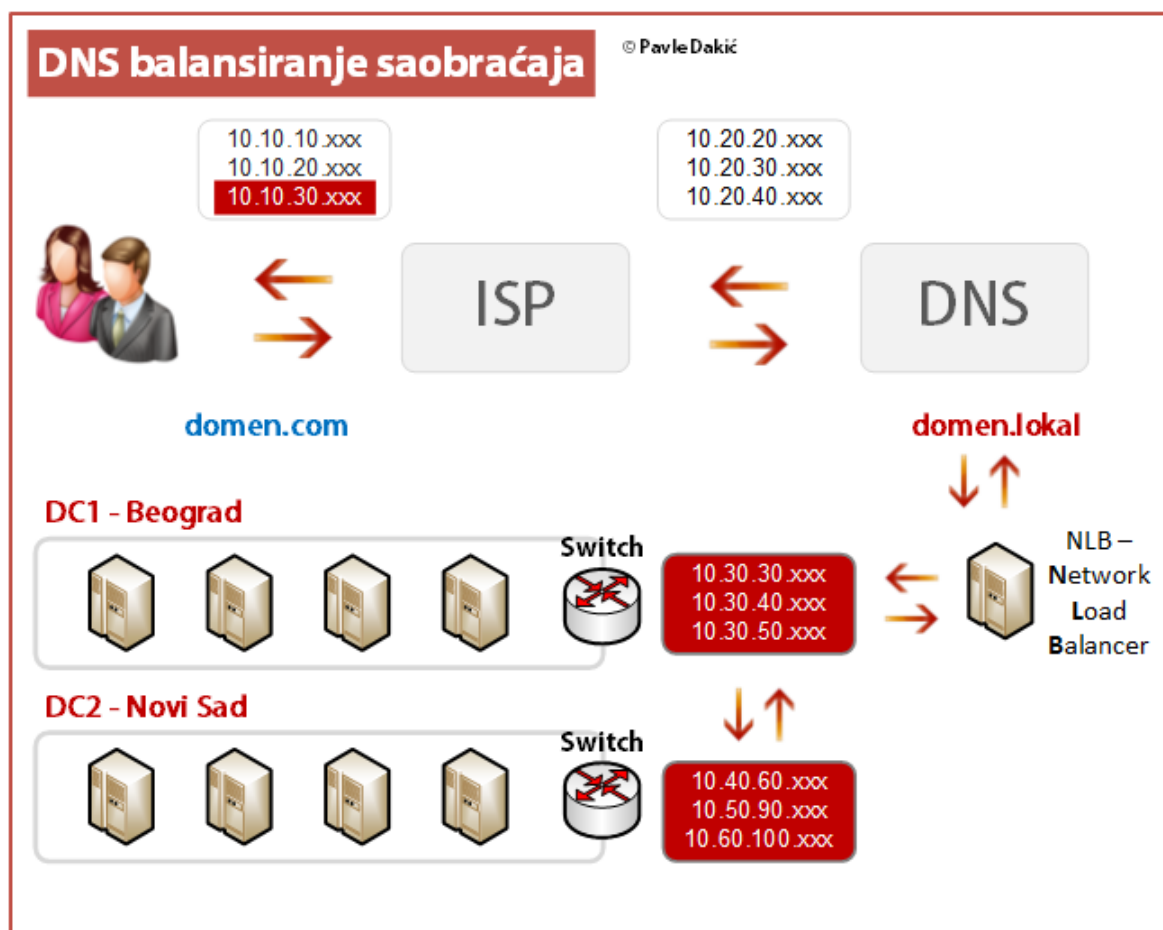
Balansiranje i kontrola mrežnog saobraćaja u velikim mrežama je veoma važna jer osigurava efikasno korišćenje resursa i zadovoljavajuće performanse sistema. U mnogim velikim računarskim mrežama balansiranje saobraćaja predstavlja izazov, zbog nedostatka globalnih informacija i koordinacije istim. Kada postoje samo lokalno dostupne informacije o trenutnom opterećenju i postojanosti određenog broja računskih nodova, teško je izvršiti centralizovani pristup i balansiranje saobraćaja (eng. Load balancing). Umesto toga je neophodno razvijanje pametnih protokola koji će garantovati brz i uravnotežen prenos poruka o različitim stanjima nodova, poštujući računarska ograničenja i mogućnosti primanja i slanja određenih informacija. Dok se ne izvrši realizacija ovih protokola HPC sistemi će zavisiti od DNS servera i njihove mogućnosti za balansiranjem i usmeravanjem saobraćaja.

4.3.1 DNS balansiranje saobraćaja

Kada korisnik ukuca ime domena, kroz niz definisanih pravila (TCP/IP protokol ili neki drugi) se vrši biranje i kontaktiranje određenih DNS name servera, koji pita DNS server za IP adresu servera koji sadrži ime domena. Ako se lista IP adresa servera se nalazi na primer na 2 servera za balansiranje saobraćaja (eng. load balancers) lociranih na dve različite lokacije, opet ako su locirani u dva različita data centra, i u slučaju da jedan od data centara postane nedostupan, klijentov računar će pokušati da ostvari komunikaciju sa prvim dostupnim DNS serverom sve dok ne dobije odgovor. Nakon uspostavljenе komunikacije između klijenta i DNS servera nastavlja se dalja komunikacija, nakon čega jedan od servera preuzima i obrađuje traženi zahtev za određenim domenom ili IP adresom. [37]

Balansiranje DNS opterećenja je praksa konfigurisanja domena da vrši distribuciju zahteva klijenata preko grupe servera. DNS služi kao "telefonski imenik" sa svim imenima računara i domena koji se postoje unutar jednog ili više DNS servera. Većina HPC aplikacija se izvršava sa više mašina, koji se nalaze iza jednog ili više servera koji vrše balansiranje saobraćaja. Da bi se sprečio mogući zastoj na mreži koristi se distribuirana sistemska arhitektura koja podrazumeva da se podaci i serveri globalno distribuiraju u različitim delovima sveta ili jedne zemlje (na primer Sombor, Novi Sad, Beograd, Niš, Zlatibor itd). Balansiranje opterećenja upotrebom DNS servera se oslanja na činjenicu da većina klijenata koriste prvu IP adresu domena koju dobiju od DNS servera (Slika 15.). U većini distribucija DNS servera se podrazumeva slanje liste IP adresa servera, u različitom redosledu prilikom opsluživanja zahteva i ova metoda se naziva round-robin. Kao rezultat toga, različiti klijenti su usmereni na različite servere, čineći sistem efikasnim i konstantno dostupnim preko grupe servera koji se nalaze unutar HPC klastera.

Nažalost ova jednostavna primena DNS balansiranja ima svoje probleme koji ograničavaju pouzdanost i efikasnost. Jer DNS server ne proverava greške i prekide, tako da uvek vraćaju isti skup IP adresa za određeni domen, čak i ako su serveri na listi nedostupni. Drugi problem nastaje kada je adresa sačuvana na lokalnom računaru, i kada dođe do promene liste IP adresa na DNS serveru nakon čega se moraju obavesti svi lokalni računari o izvršenim izmenama zbog čega svaka dodeljena IP adresa ima određeni životni vek trajanja pod nazivom TTL ili time to live, sa ciljem da klijenti imaju uvek ažurnu listu aktivnih servera. Kratki životni vek poboljšava preciznost i omogućava keširanje (eng. caching), ali zahteva konstantno ažuriranje i praćenje stanja aktivnih servera. [48]



Slika 40. DNS Balansiranje saobraćaja, data centri Beograd, Novi Sad.

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

Ako bismo želeli da podesimo naše servera, tako da DNS zahtevi klijenata (u ovom slučaju naziv/i domena) koriste round robin rotaciju potrebno je da dodamo više A zapisa (Tabela 16.) unutar Autoritativnog DNS servera (eng. Authoritative DNS name server). Da bi kasnije bilo moguće pristupanje imenu domena (www.domen.com) preko bilo kog DNS servera neophodno je podesiti prosleđivanje zahteva na druge servere, TTL vrednosti čuvanja zapisa na lokalnom računaru i IP adresa servera koji sadrži DNS zonu. Prikazana vrednost (60) u Tabeli 16. predstavlja vreme čuvanja nekeširanog zapisa u trajanju od 60 sekundi na lokalnom računaru i koristi se za posredne servere koji ne podržavaju upotrebu Round Robin tehnologije. [48]

Tabela 16. Balansiranje saobraćaja domena upotrebom više servera.

domen.com	60	IN	A	10.30.30.1
domen.com	60	IN	A	10.40.60.1
domen.com	60	IN	A	10.50.90.1
domen.com	120	IN	A	10.50.90.2

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

5 Operativni sistemi HPC

Operativni sistemi korisnicima predstavljaju samo platformu za izvršavanje određenih zadataka zbog čega korisnici žele operativni sistem koji poseduje mogućnost primene na HPC sistemu, za potrebe razvoja njihovog radnog okruženja upotrebom bilo kojeg operativnog sistema Windows, Mac, Linux itd. Unutar svake kompanije koja se bavi ozbiljno određenom oblašću, postoje korisnici koji intenzivno koriste određene programe, koji se izvršavaju samo pod određenim operativnim sistemom. Izbor pravog operativnog sistema najviše zavisi od broja korisnika i programa koji se koriste u radnom okruženju. Operativni sistemi koji poseduju mogućnost realizacije HPC sistema su Windows i Linux.

U slučaju da se programi pokreću pod **Windows** operativnih sistemom, moraju se uzeti u obzir dodatni troškovi vezani za licence i softver koji se koristi. Windows poseduje jedan operativni sistem Windows HPC server, verzije 2008 i 2012 koji omogućavaju centralizovanu integraciju unutar HPC okruženja zajedno sa Aktivnim direktorijumom (eng. Active Directory) i svim ostalim windows servisima. Međutim problem kod Windows HPC operativnog sistema je u tome što je microsoft smanjio veličinu svog HPC razvojnog tima smanjujući broj implementacija HPC sistema, upotrebom Windows HPC operativnog sistema. Windows HPC nije naročito lagan po pitanju veličine operativnog sistema, tako da jedan nod zauzima minimalno 10 GB prostora na hard disku. Kao glavni razlozi za nekorišćenjem Windows HPC rešenja koji su naveli nezadovoljni korisnici je bilo: potrebno vreme za instalaciju, kompleksnost konfiguracije, problemi prilikom dodavanja novog računskog noda itd.

Za razliku od Windows operativnog sistema **Linux** je lagan i zauzima veoma mali prostor na hard disku (~512 - 1536 MB) u zavisnosti od vrste računskog noda. Većina Linux distribucija sadrži zajednički softver za povezivanje nodova, web servise za nadgledanje i kontrolu saobraćaja omogućavajući povezivanje različitih Linux distribucija unutar HPC sistema. Linux je potpuno prilagodljiv i jednostavan za upravljanje i optimizaciju resursa. Ovo ga čini najpogodniji sistem za izgradnju i osnivanje velikih sistema koji sadrže veliki procesora. Pored toga većina naučnih programa i softvera za kompajliranje koda su lako dostupni za Linux kroz paket menadžere unutar operativnog sistema, ili direktno putem Interneta. Izbor distribucije najviše zavisi od namene HPC sistema, i većinom izbor distribucije je stav ličnog opredeljenja. [4]

Kada je reč o povezivanju i podršci za gore navedene protokole (InfiBand, Interconnect, MPI, OpenMPI itd.) većina HPC operativnih sistema sadrži zadnje verzije komunikacionih protokola. Na sledećoj listi se nalaze operativni sistemi (Linux, Windows) koji poseduju najčešće i najkorišćenije implementacije HPC sistema:

1. Linux
 - 1.1. RocksClusters
 - 1.2. Red Hat Enterprise Linux
 - 1.3. CentOS
 - 1.4. SUSE Linux Enterprise Server
2. Windows
 - 2.1. Windows server 2008 R2 HPC
 - 2.2. Windows server 2012 R2 HPC

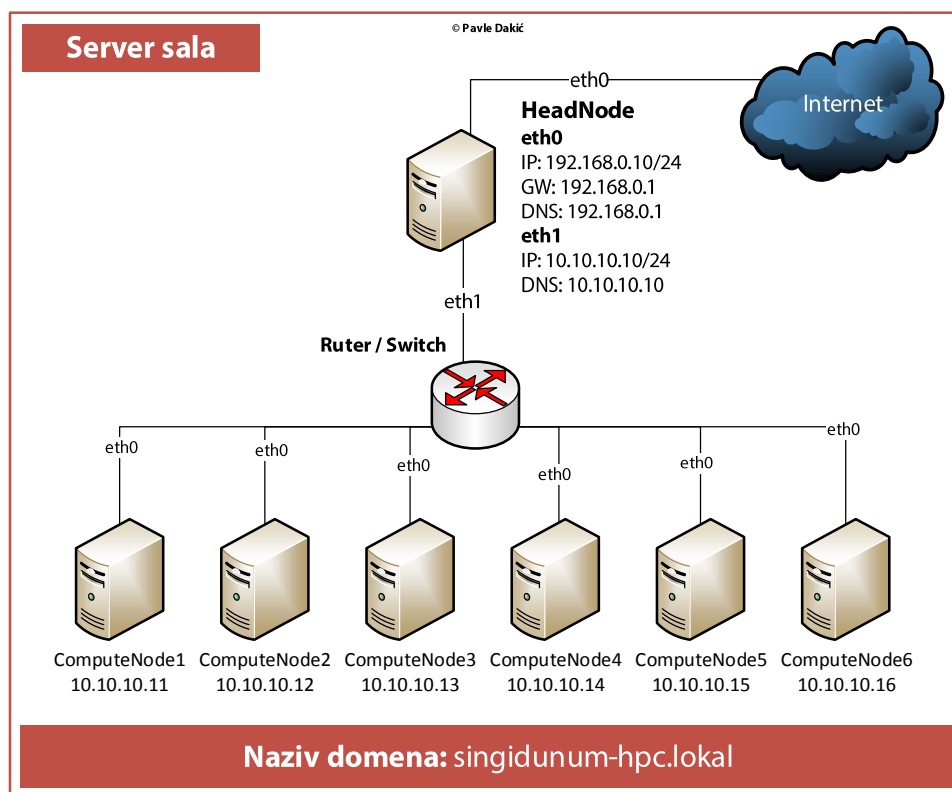
6 Laboratorijsko okruženje

Koristeći laboratorijsko okruženje za testiranje i razvoj HPC sistema, možemo da izvršimo konfiguraciju i testiranje određenog operativnog sistema. Testiranjem unutar virtuelnog okruženja možemo uraditi sve ono što bi kasnije radili prilikom fizičkog razvoja sistema visokih performansi, smanjujući nepotrebne troškove prilikom kupovine hardvera i softvera. Migracija uživo je jedna od glavnih prednosti virtualizacije i laboratorijskog okruženja, jer omogućava migraciju sa virtuelne mašine na fizički host bez ikakvih zastoja u samom radu sistema. Kako bi se mogla izvršiti migracija sa virtuelne mašine na fizički host, potrebna je dostupnost NFS ili iSCSI uređaja za čuvanje podataka koji je dostupan na mreži ili unutar fizičkog računara, na kojem se nalazi virtualna mašina.

Za realizaciju i testiranje korišćićemo **VMware Workstation** i **RocksClusters Linux** operativni sistem (<http://www.rocksclusters.org/>) sa konfiguracijom prikazanoj u Tabela 17. gde ćemo izvršiti osnovnu konfiguraciju i dodavanje nekoliko računskih nodova, koji će činiti naš HPC sistem (Slika 41.).

Tabela 17. Specifikacija laboratorijskog okruženja.

Naziv domena	singidunum-hpc.lokal	
Ethernet 0 (Private)	192.168.0.0/24	
Ethernet 2 (Public)	10.10.10.0/16	
HeadNode	Ethernet 0 (Private)	Ethernet 2 (Public) Internet
	10.10.10.10/16	192.168.0.10/24
ComputeNode1	10.10.10.11	
ComputeNode2	10.10.10.12	
ComputeNode3	10.10.10.13	
ComputeNodeX	10.10.10.X	

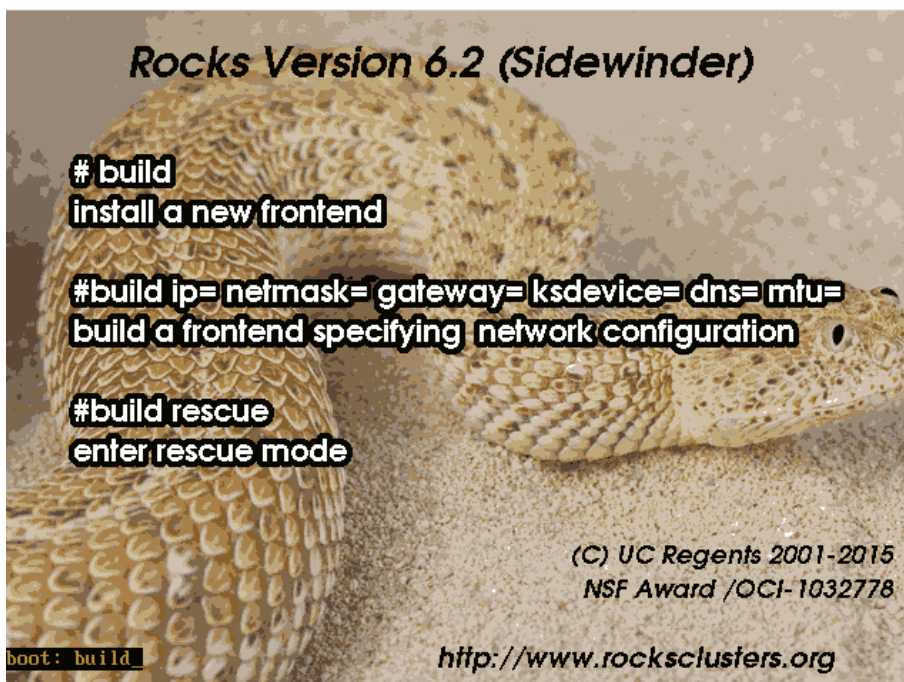


Slika 41. Lab. - Šema laboratorijskog okruženja

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.1 Izgradnja HPC klastera - Build

Kako bi napravili novi HPC klaster, neophodno je da u donjem levom uglu unesemo komandu: **build** nakon čega se pokreće mod za instalaciju klastera i izbor aplikacija/rola (Slika 43.).

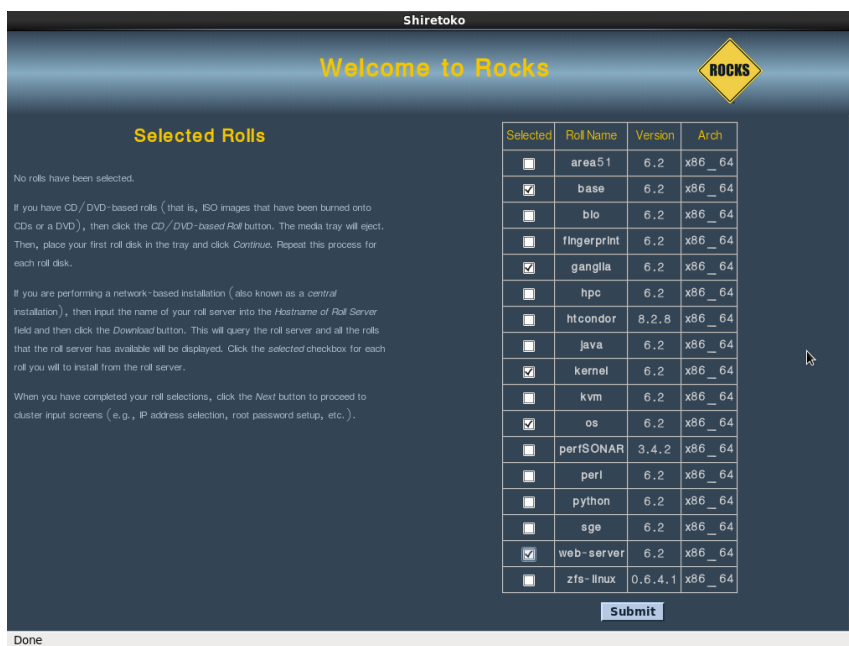


Slika 42. Lab. - RocksCluster v6.2 (Sidewinder)

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

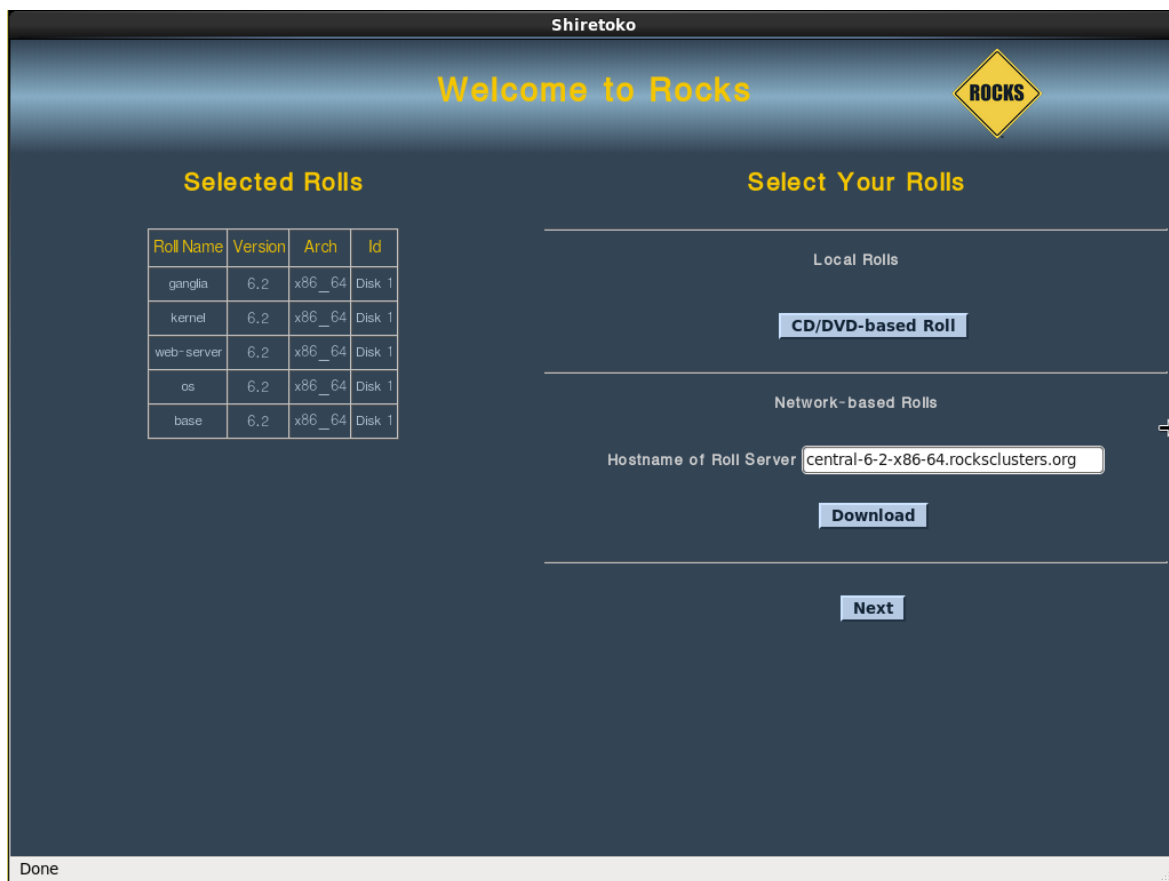
6.2 Izbor odgovarajućih servisa - Rolls

Rock Cluster poseduje mogućnost izbora servisa koji će se pokretati u vidu različitih paketa kao što su ganglia za nadzor saobraćaja, java i itd. Servisi koje smo izabrali su: **base**, **ganglia**, **kernel**, **os**, **web-server**.



Slika 43. Lab. - Izbor odgovarajućih servisa – Rolls

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.



Slika 44. Lab. - Izabrani odgovarajući servisi – Rolls

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.3 Informacije o klasteru

Instalacija HPC klastera zahteva unos određenih informacija, puno ime računara/servera, domena, ime klastera, sertifikaciono telo/organizacija, lokalitet, država, grad, kontakt, url (web sajt), GPS koordinate (Slika 45.).

The screenshot shows the 'Cluster Information' form. It includes a 'Help' sidebar on the left and a main form area on the right. The form fields are as follows:

Field	Value
Fully-Qualified Host Name	singidunum-hpc.lokal
Cluster Name	Singidunum-HPC-klaster
Certificate Organization	Univerzitet Singidunum
Certificate Locality	Danijelova 32, Beograd
Certificate State	Srbija
Certificate Country	RS
Contact	admin@singidunum-hpc.lokal
URL	http://www.singidunum-hpc.lokal/
Latitude / Longitude	N44.8167 E20.4667

At the bottom of the form are 'Back' and 'Next' buttons.

Slika 45. Lab. - Osnovne informacije o Klasteru.

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.4 Podešavanje mreže

Nakon unesenih osnovnih informacija o klasteru, vrši se podešavanje mrežnih kartica **eth1** i **eth0** (Slika 46,47.).

Shiretoko

Welcome to Rocks

ROCKS

Help

Ethernet Configuration for Public Network

Public Network Interface:

This is the Ethernet network that physically connects your frontend to the outside world.

Public Network Interface: eth1

IP address: 192.168.0.10

Netmask: 255.255.255.0

mtu: 1500

Back Next

Slika 46. Lab. - Podešavanje Javne mreže na kartici eth1.

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

Shiretoko

Welcome to Rocks

ROCKS

Help

Ethernet Configuration for Private Network

Private Network Interface:

This is the Ethernet network that physically connects your frontend to compute nodes.

Private Network Interface: eth0

IP address: 10.10.10.10

Netmask: 255.255.0.0

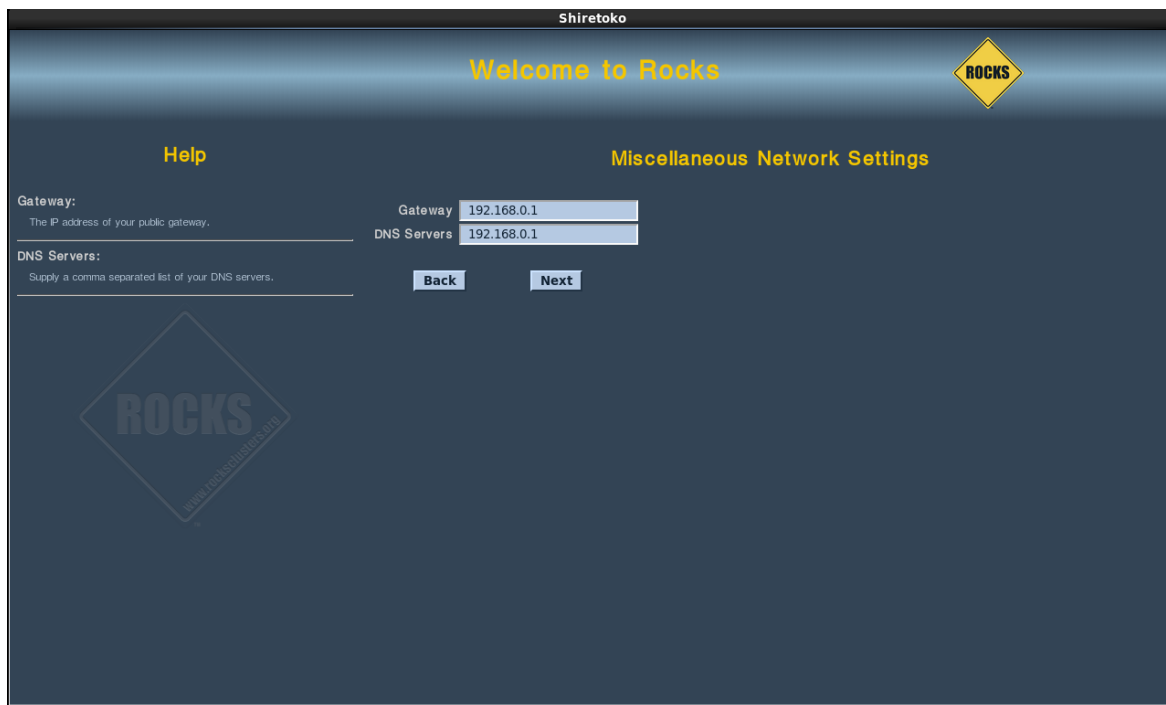
Back Next

Slika 47. Lab. - Podešavanje Privatne mreže na kartici eth0

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.5 Gateway, DNS server

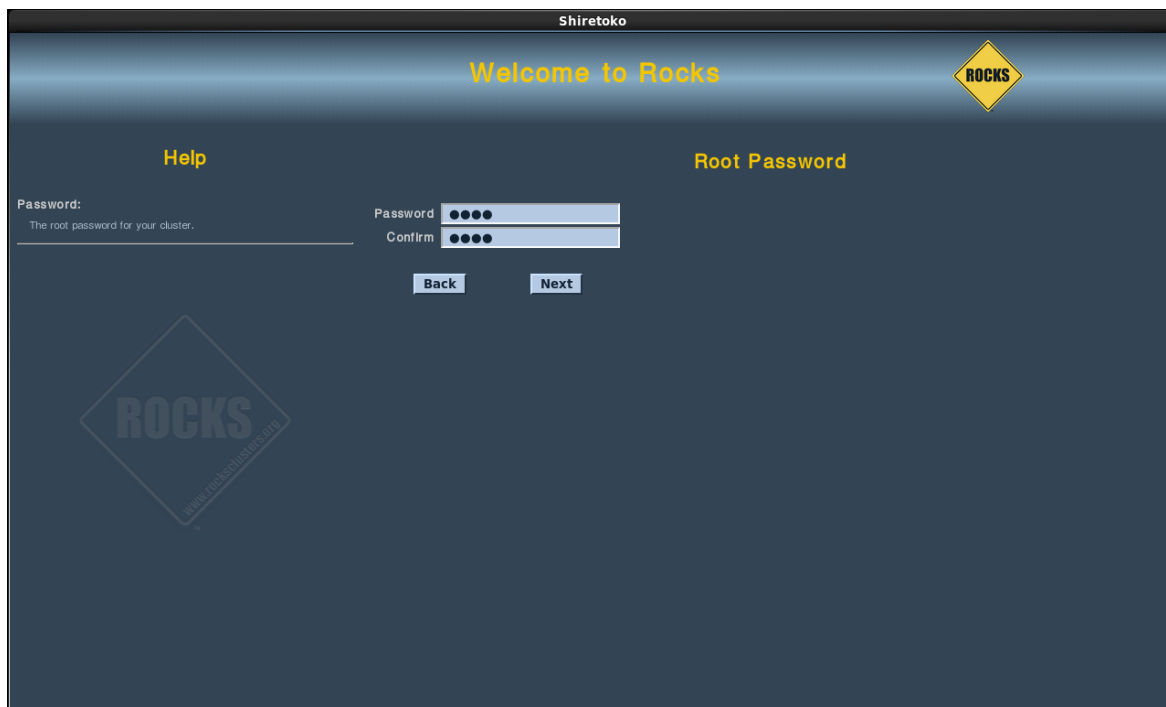
Podešavaju za Gateway i DNS treba da budu iz Public mreže, koja izlazi na internet, a kasnije tokom daljeg razvoja HPC sistema bi trebalo staviti lokalni DNS servera koji će vršiti balansiranje saobraćaja.



Slika 48. Lab. - Podešavanje Privatne mreže na kartici eth0

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.6 Root Password

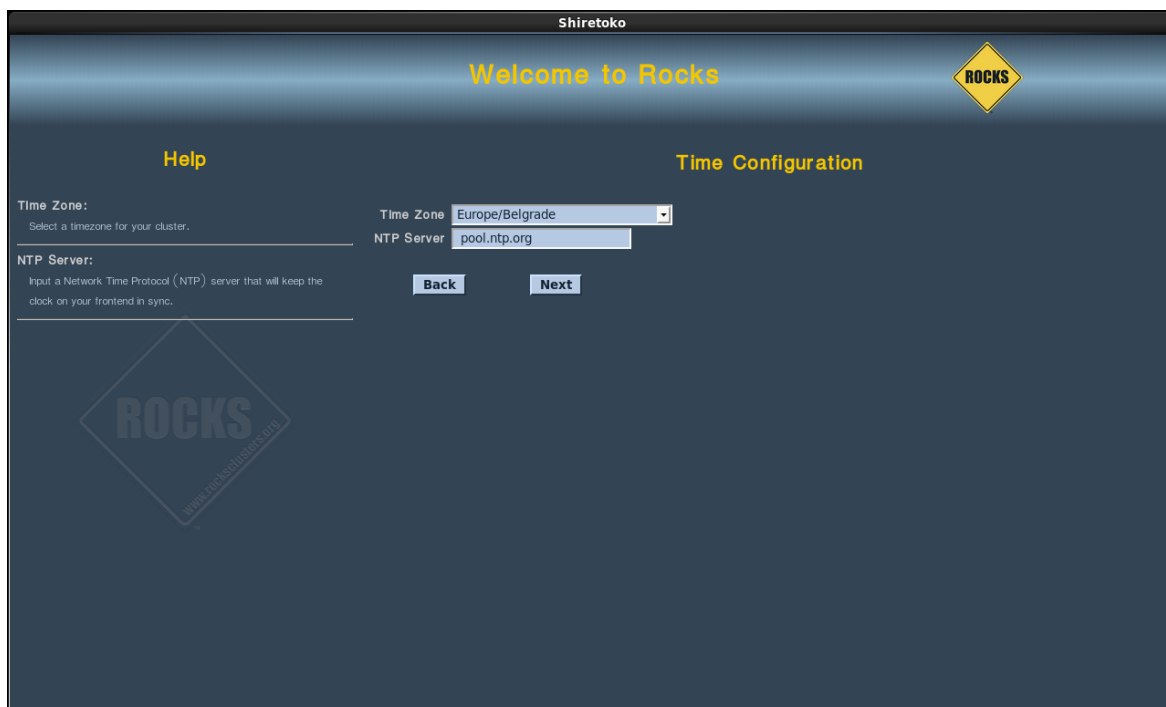


Slika 49. Lab. - Podešavanje password-a za nalog root.

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.7 Vremenska zona, NTP server

Podešavanje vremenske zone i servera za sinhronizaciju, kako bi se smanjile greške validnosti vremena na file sistemu.

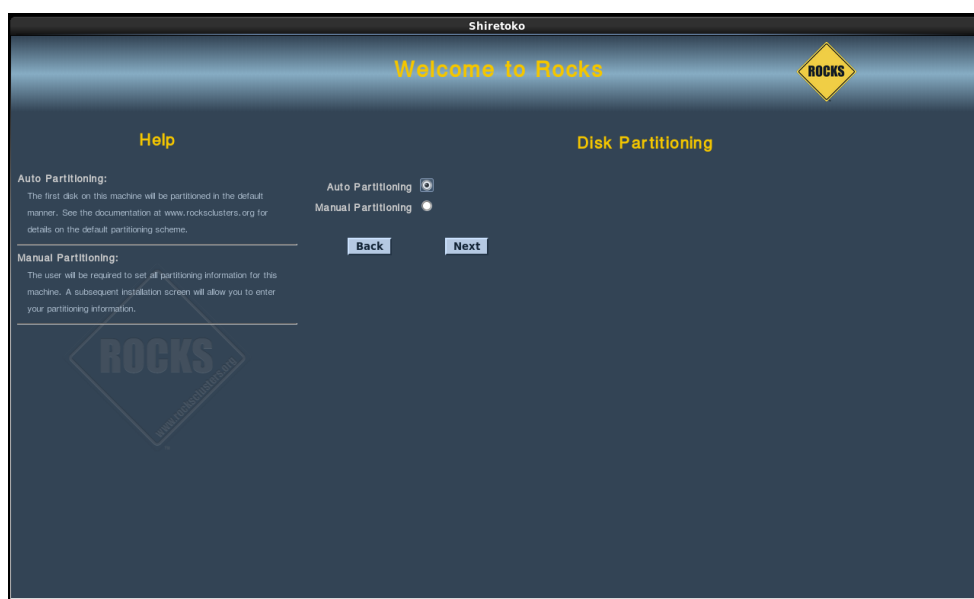


Slika 50. Lab. - Podešavanje vremenske zone, NTP server.

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.8 Automatsko Particionisanje i formatiranje hard diska

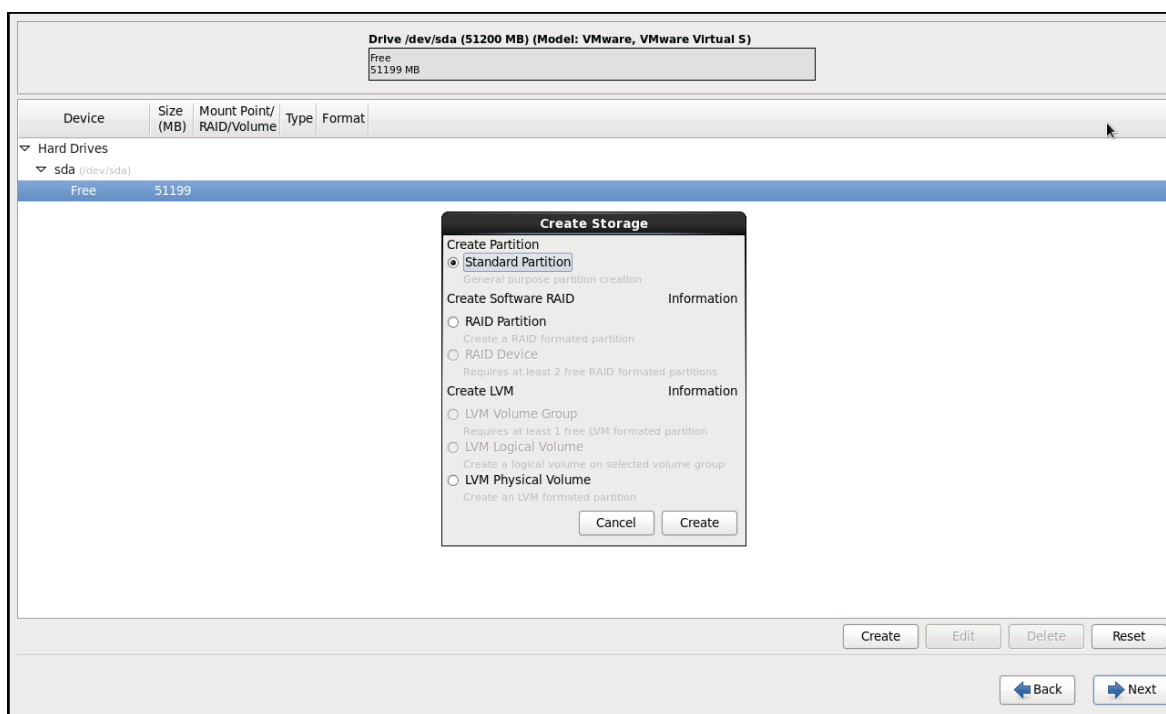
Particionisanje diska može se realizovati na dva načina: **Auto partitioning** (Slika 51.), gde sistem automatski priprema i vrši formatiranje hard diska i drugi način **Manual partitioning** (Slika 52,53.) koristimo kada na sistemu posedujemo više od jednog hard diska, kako ne bi došlo do gubitka podataka.



Slika 51. Lab. - Particionisanje diska.

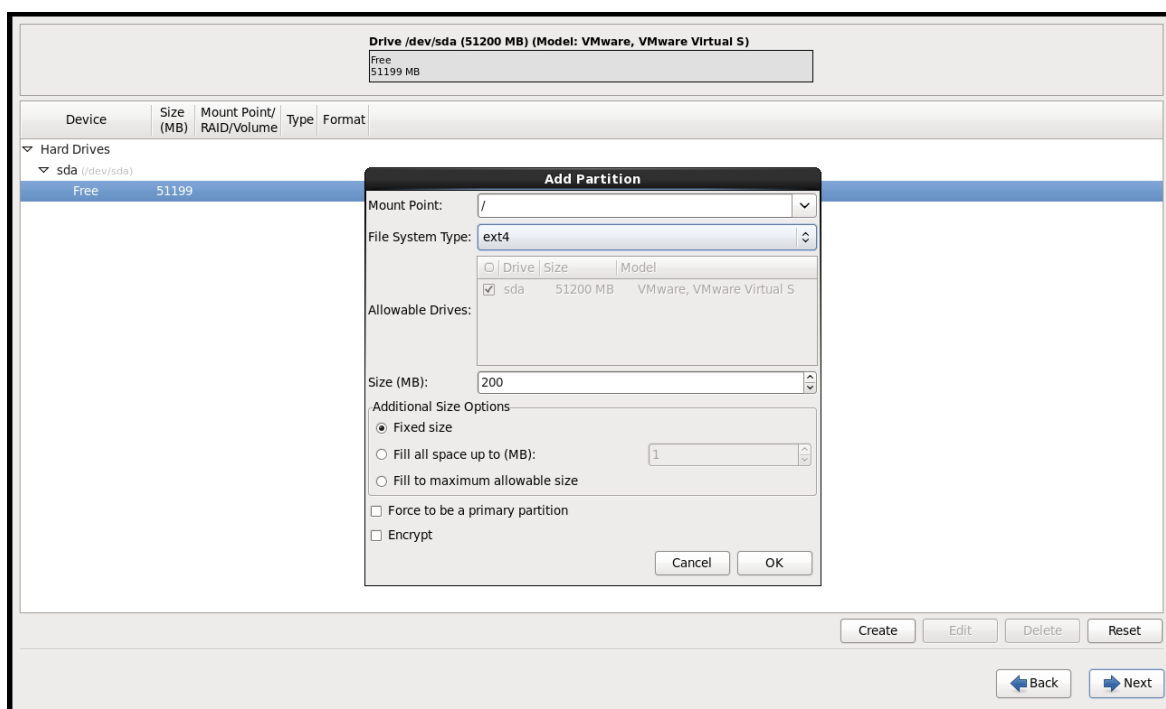
Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.9 Ručno Particionisanje i formatiranje hard diska



Slika 52. Lab. – Ručno Particionisanje diska.

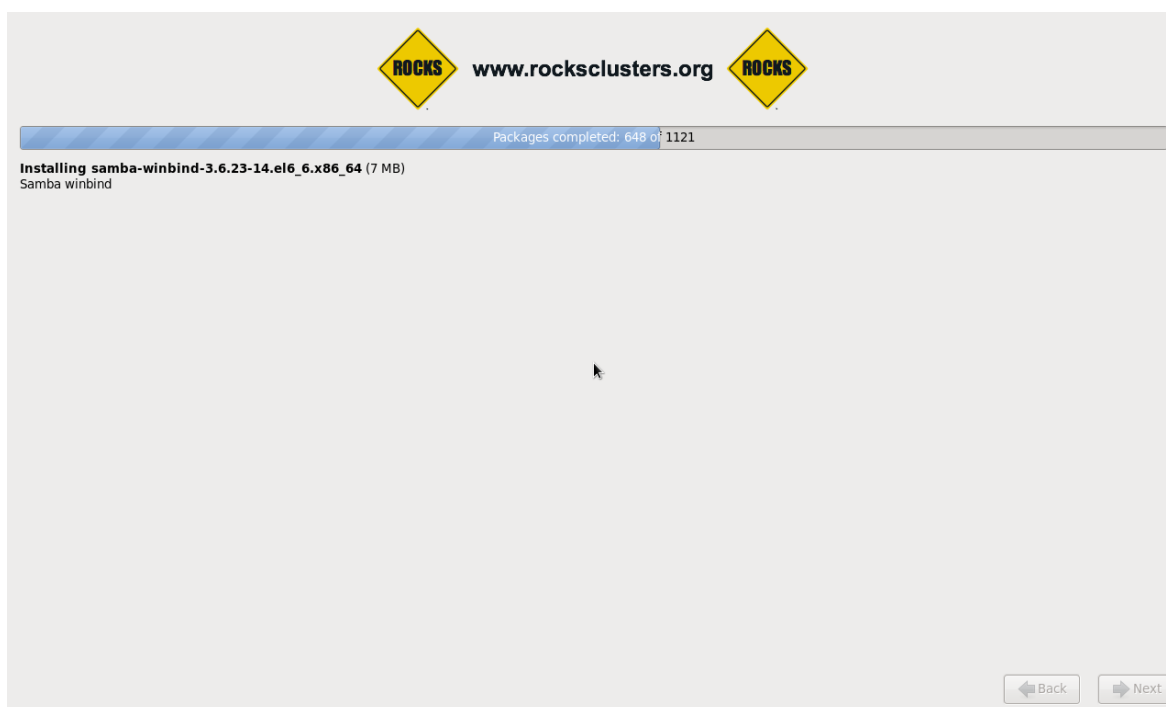
Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.



Slika 53. Lab. – Ručno Particionisanje diska, dodavanje particije.

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.10 Instalacija klastera

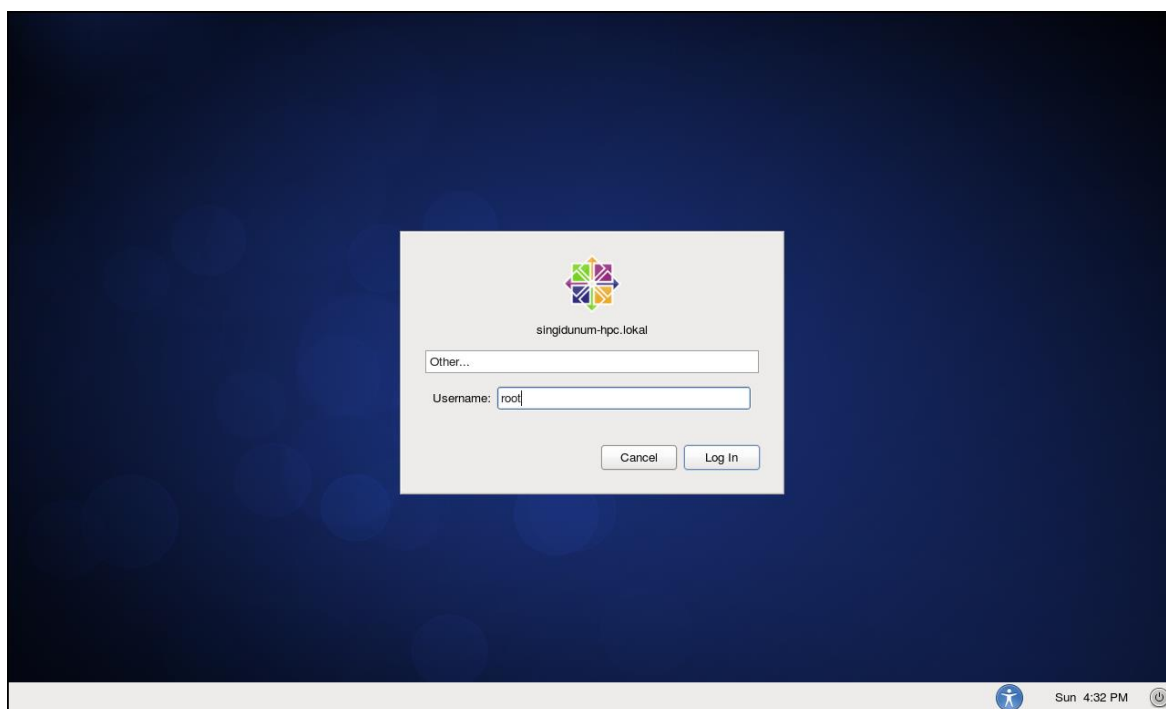


Slika 54. Lab. - Instalacija klastera.

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.11 Logovanje na sistem

Logovanje na sistem se vrši unosom username-a i password-a koji smo prethodno definisali.

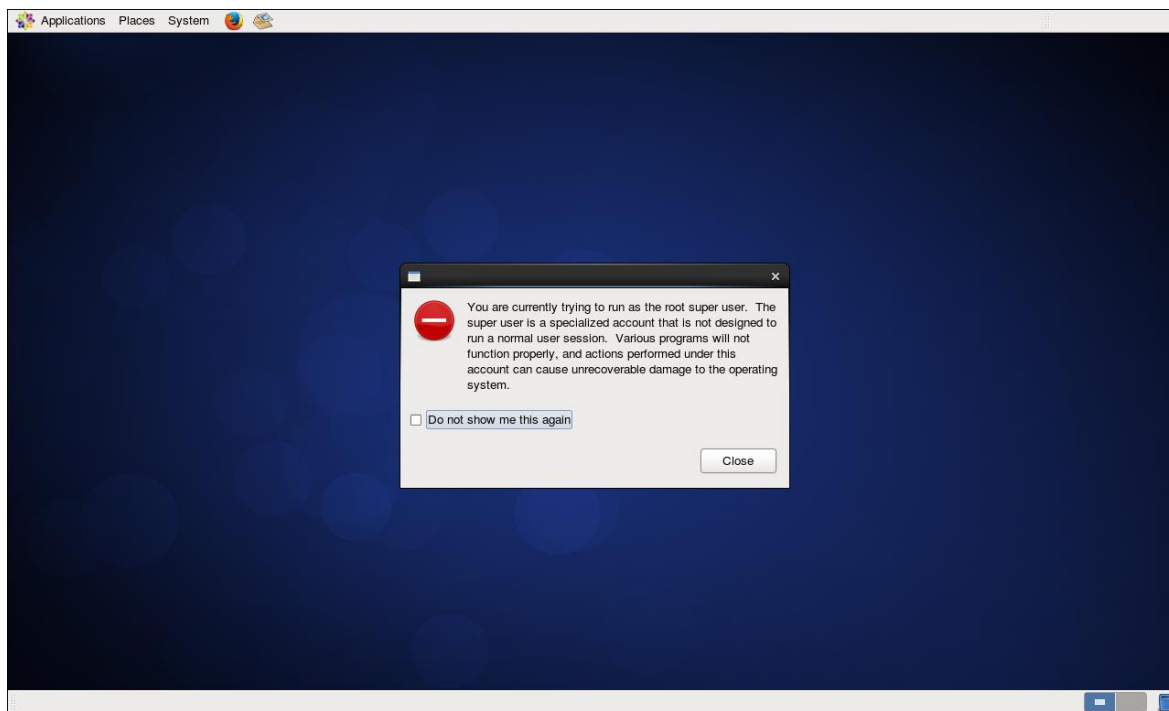


Slika 55. Lab. - Logovanje na sistem

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.12 Logovanje na sistem, upozorenje

Nakon logovanja na sistem dobićemo poruku sa upozorenjem da koristimo nalog koji poseduje prava super korisnika, i da nije preporučljivo da ga pokrećemo tokom regularnih sesija. Kako bi korisnici mogli da koriste sistem sa bilo kojeg računara, neophodno je napraviti korisničke naloge (`# useradd singi`) i izvršiti sinhronizaciju sa svim nodovima unutar klastera (`# Rocks sync users`).

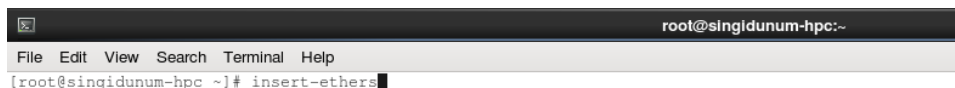


Slika 56. Lab. - Logovanje na sistem

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.13 Dodavanje novih računskih nodova

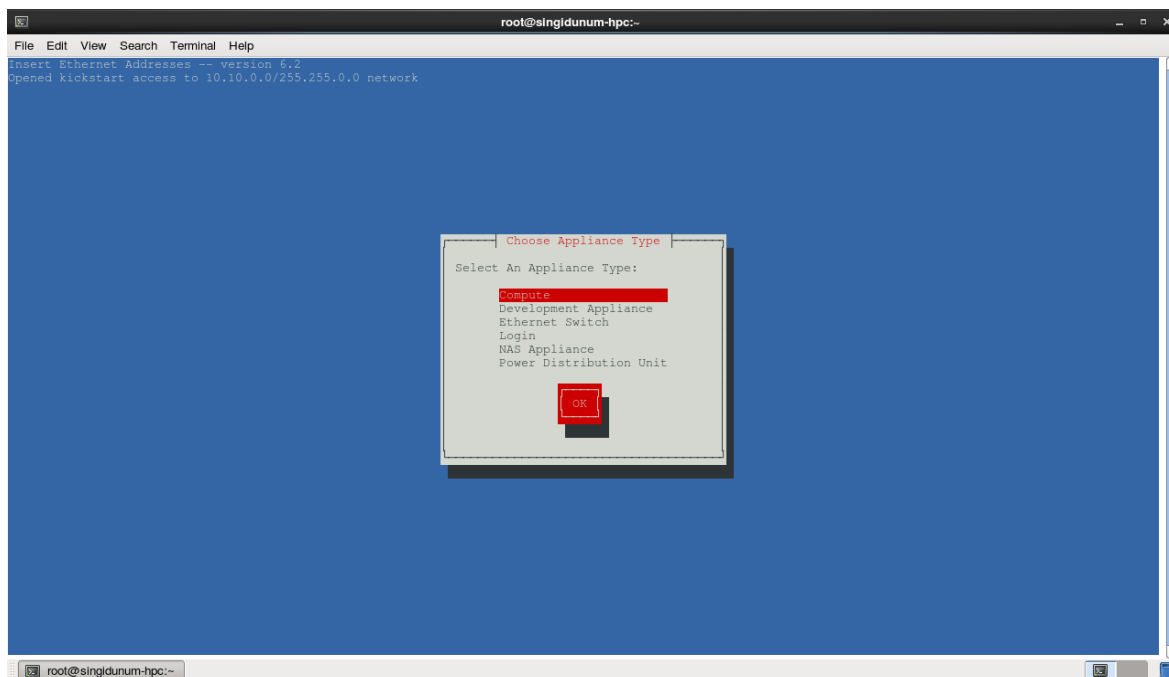
Proces dodavanja novih računskih nodova ili drugih operacija se realizuje pokretanjem terminala i unosom komande: **insert-ethers** nakon čega se prikazuje prozor (Slika 57.) za izbor određene stavke. Za naše potrebe koristićemo prvu opciju **Compute** (Slika 58.).



Slika 57. Lab. - Dodavanje novog noda

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.14 Dodavanje novih računskih nodova, opcija Compute

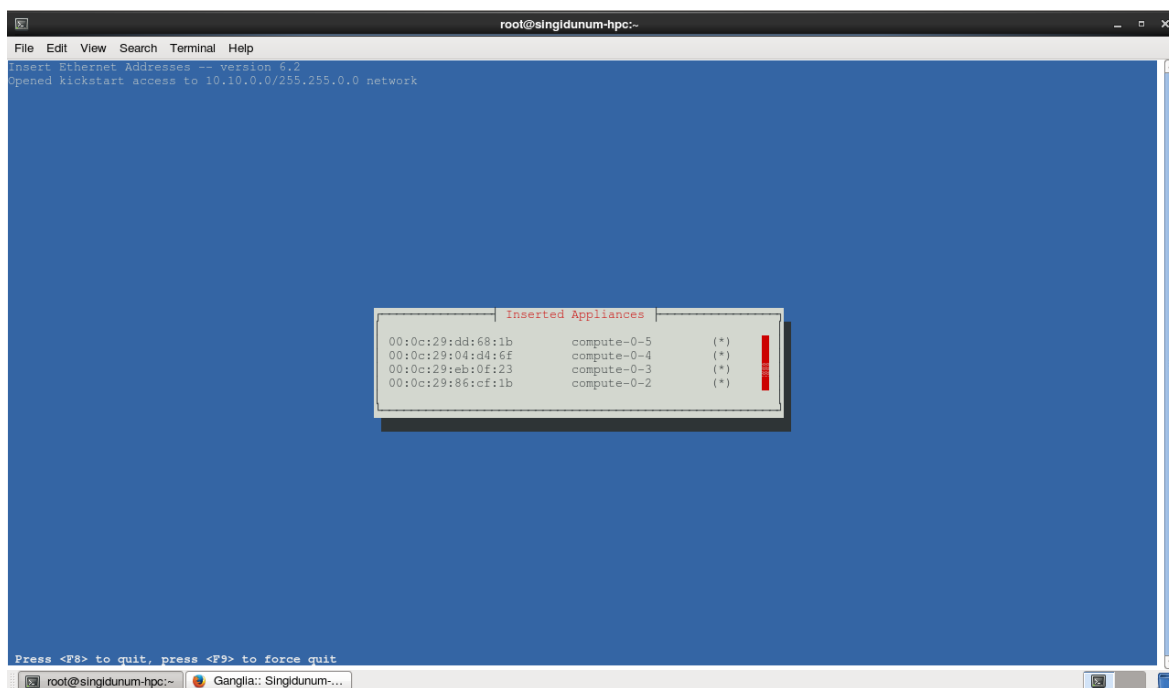


Slika 58. Lab. - Dodavanje novog noda

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

6.15 Dodavanje novih računskih nodova, ostvarivanje komunikacije

Nakon što smo izabrali opciju **Compute** dobijamo prozor koji prikazuje sve nodove koji su uspešno ostvarili komunikaciju sa našim HPC klasterom (Slika 59.).

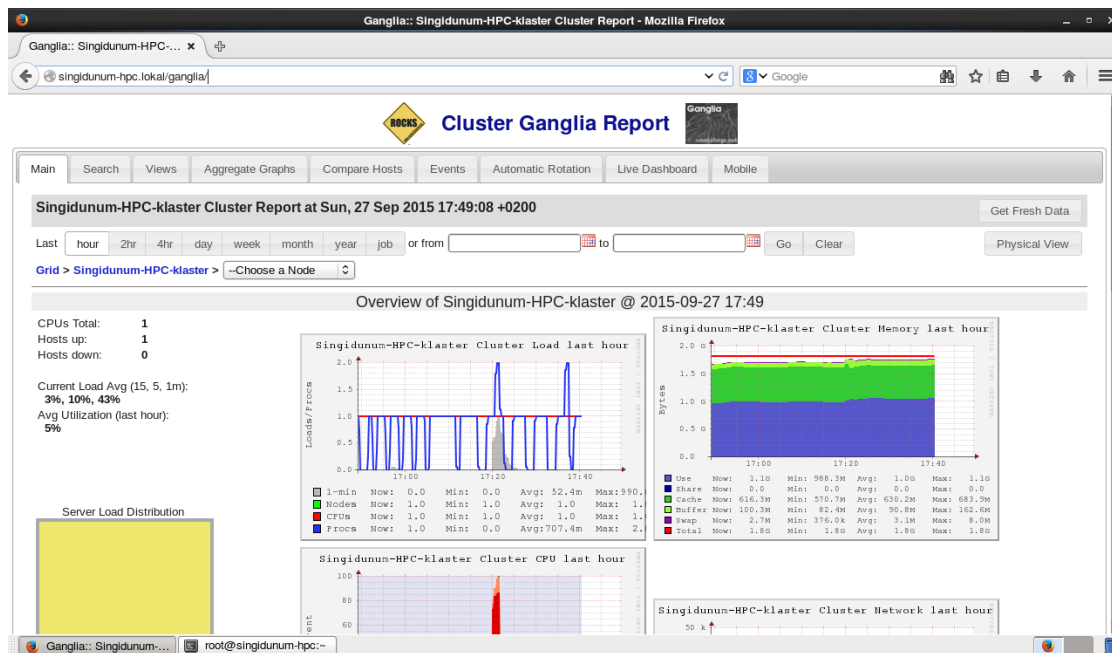


Slika 59. Lab. - Dodavanje nodova

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

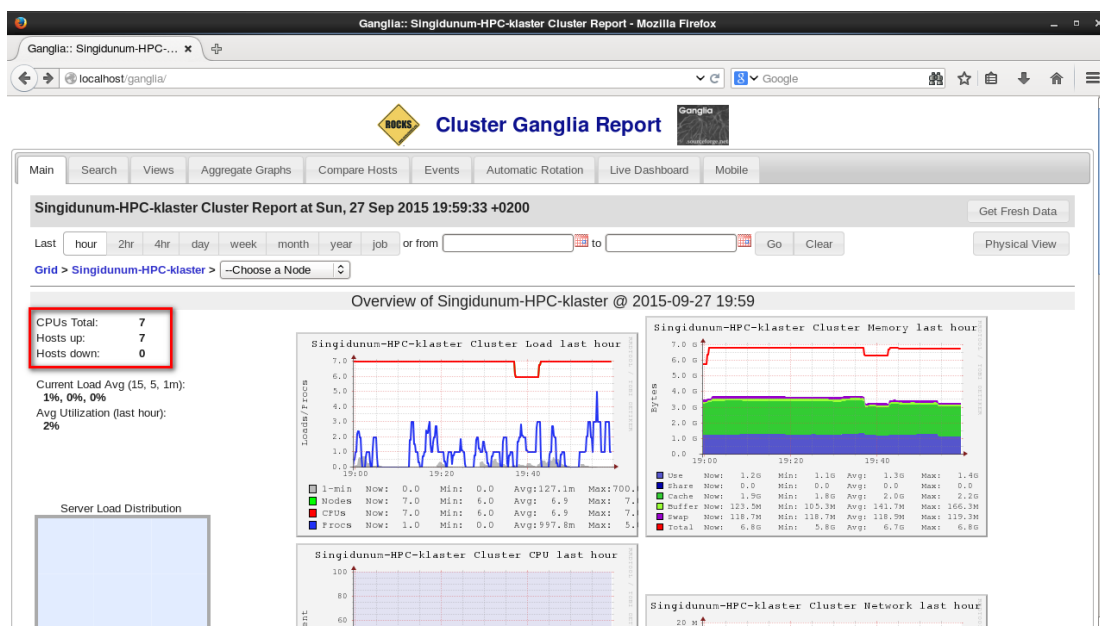
6.16 Ganglia

Ganglija je skalabilni distributivni monitoring sistem namenjen za računarsvo visokih performansi, kao što su klasteri i mreže. Zasniva se na hijerarhijskom dizajnu usmerenom na federaciju klastera. Ganglija usklađuje široku primenu tehnologija kao što su MPI, Nvidia Cuda, XML itd. pomoću kojih vrši predstavljanje podataka koji se nalaze na svakom od nodova. Za dobijanje podataka u realnom vremenu koristi određene algoritme koji vrše sakupljanje i slanja podataka HeadNodu. Sama implementacija je dosta robusna i dobro razvijena i sadrži široku paletu podrške za različite operative sisteme, procesore, RAM memorije, komponente i može da podrži nadgledanje do **2.000** računskih nodova.



Slika 60. Lab. – Osnovna konfiguracija HPC sistema sa HeadNodom.

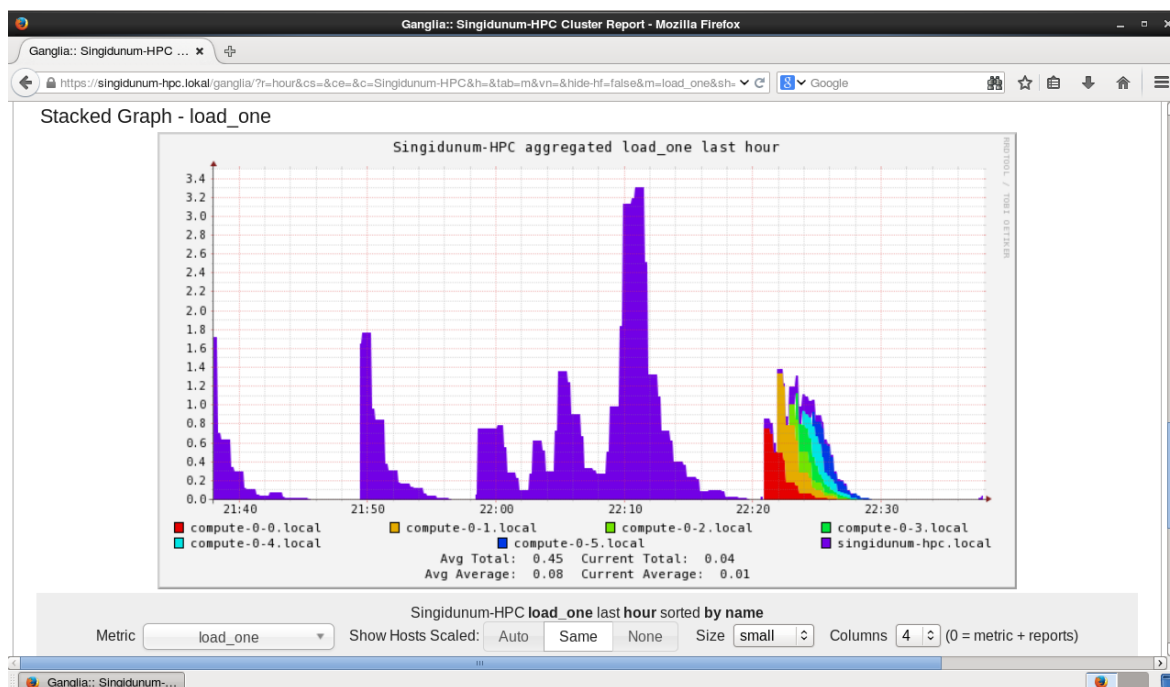
Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.



Slika 61. Lab. – Osnovna konfiguracija HPC sistema HeadNod i 6 nodova.

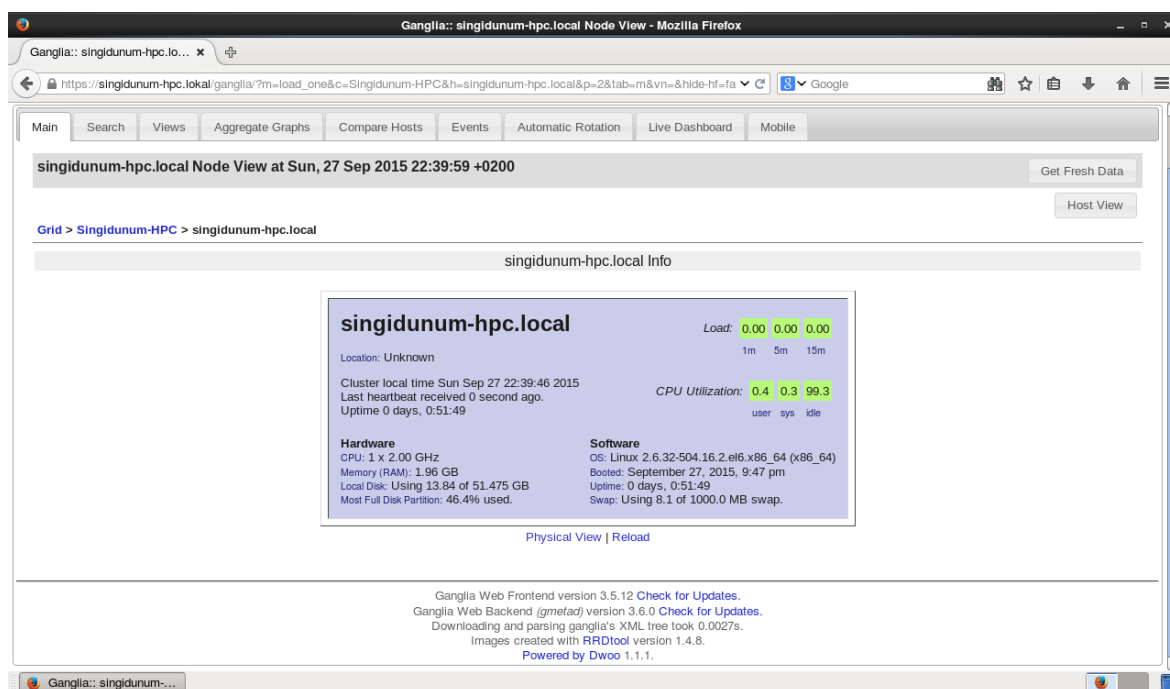
Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

Na prikazanoj slici (Slika 62.) može se videti trenutno opterećenje nodova, pri čemu svaki nod je označen drugom bojom.



Slika 62. Lab. – Pregled opterećenja po nodovima.

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

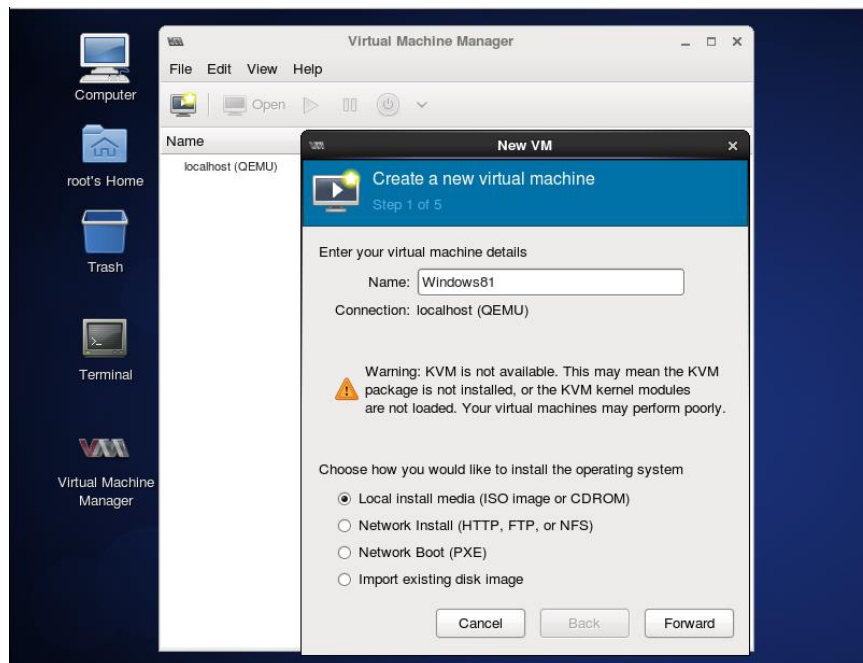


Slika 63. Lab. – Pregled hardverskih mogućnosti singidunum-hpc.local.

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

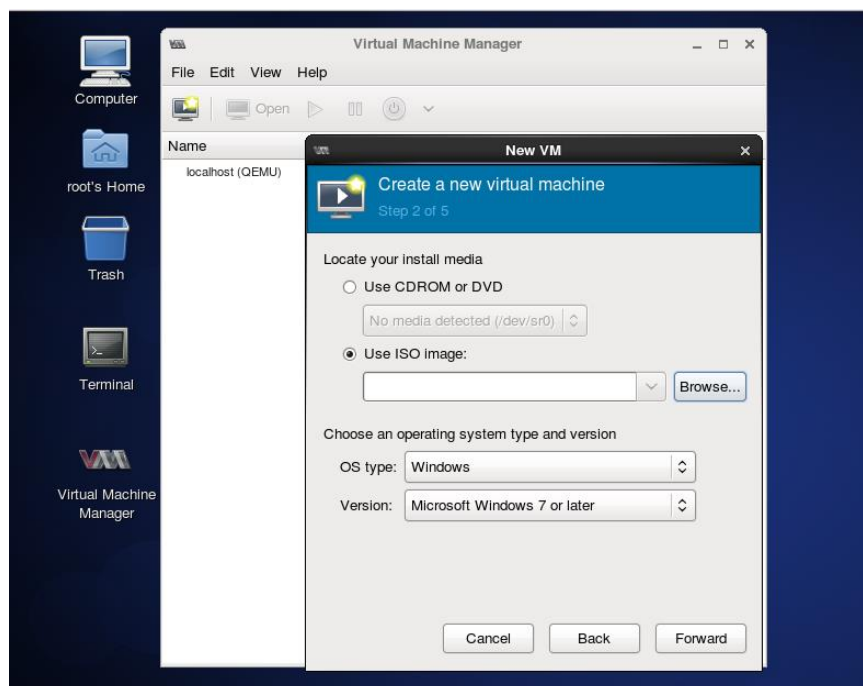
6.17 Virtualizacija sistema

Zahvaljujući postojanosti virtualizacije veoma lako se rešava problem operativnih sistema koji mogu da koriste HPC klaster. Da bi bilo moguće korišćenje resursa svih povezanih računskih nodova, sa svim njihovim povezanim komponentama, na glavnom nodu (HeadNodu) je potrebno izvršiti konfiguraciju i instalaciju potrebnog sistema (Slika 64,65.).



Slika 64. Lab. – Virtualizacija sistema.

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.



Slika 65. Lab. – Virtualizacija sistema #2.

Izvor: sopstvena izrada, na osnovu podataka prikupljenih unutar istraživačkog rada.

Zaključak

U ovom master radu su predstavljena trenutna hardverska rešenja i pregled njihovih mogućnosti. Dobijeni su rezultati analize kompanija koje se bave realizacijom ovih sistema.

Super računari su poput "motora" automobila, vode nas u susret otkrićima koja imaju značajan uticaj na naš svakodnevni život ostvarujući neograničenu ljudsku maštu, upotrebom računarske snage HPC sistema, a sa druge strane HPC sistemi poseduju nezasitnu žeđi za prostorom, hlađenjem, električnom energijom, brzom komunikacijom, osobljem itd. Vreme promena u IT industriji, zahtevaće promenu kursa i načina proizvodnje servera. Razlog za to je rast svetske populacije i dostizanje broja od 10 milijardi ljudi, što će dovesti do znatnog povećanja količine podataka, koji će se razmenjivati i čuvati na pametnim telefonima (eng. Smarthphone), računarima, serverima i drugim uređajima za skladištenje podataka. Mobilni telefoni i drugi uređaji koji imaju mogućnost internet komunikacije, su jedan od glavnih generatora današnjeg sadržaja/informacija u vidu fotografija, video materijala, zvuka i drugih multimedijalnih sadržaja, zbog čega će u budućnosti svaki korisnik generisati određenu količinu podataka, koju će biti potrebno sačuvati, obraditi i učiniti dostupnim za što kraći vremenski period. Jedna od prednosti HPC je sposobnost postizanja velike brzine izračunavanja, koristeći postojeće performanse CPU/GPU unutar njihovih granica. Klasteri za izračunavanje su postali najtraženije hardversko rešenje, zbog svoje produktivnosti i fleksibilnosti simulacija i samim tim postali sastavni deo svake naučne oblasti, koja se ozbiljno bavi naukom maksimizirajući produktivnost i podstičući kreativnost samih korisnika prilikom rešavanja i definisanja problema za rešavanje.

Evolucija serverskih tehnologija i sistema visokih performansi je omogućila dramatično smanjivanje troškova pružajući za hiljadu puta veću računarsku snagu u samo nekoliko godina, omogućavajući da više kompanija može da iskoristiti moć super računarstva u obliku HPC klastera. Razlog za to je trenutna neodrživost postojećih data centara i njihovih zahteva za proširenjem u obliku prostora, što svakako zahteva dodatna ulaganja. Što više ljudi može da dobije u svoje ruke ovu vrstu kapaciteta više otkrića, inovacija i mogućnosti ćemo moći da vidimo kao rezultat. U stvari, čak i posedovanje delića HPC mogućnosti u rukama obrazovnih ljudi u Srbiji je generalno dovoljno za rešavanje mnogih problema, koje možda sami i ne možemo ni da zamislamo da mogu da se reše. Većina super računara, kao i računarskih oblaka, su namenjena za više korisnički rad (eng. multi-tenant) s' razlogom da jedan korisnik koristi određene delove celokupnog sistem, omogućavajući fleksibilan rad više korisnika u bilo kom trenutku. Za održivost poslovanja rukovodioci će morati da osmišljavaju i grade svoje poslovanje na osnovu tehnologija koje će prosperirati u budućnosti gde će novo nastali koncepti poput veštačke inteligencije, 3D štampe, pametnih uređaja itd. biti sastavni deo svake kompanije. Drugim rečima kompanije će morati da poseduju svoju "digitalni biljku" koja će moći da održi korak sa razvojem budućih tehnologija. Računarska snaga će biti jeftina i svima dostupna, pri čemu nećemo primećivati razliku između PC i računara visokih performansi.

Na osnovu izvršenog istraživanja, u budućnosti ćemo imati priliku da vidimo različite uređaje povezane na računarsku mrežu koji će posedovati i koristiti hardverske HPC uređaje, poput automobila, bele tehnike, vozova, televizora itd. Takođe veći broj profesionalnih korisnika će posedovati neki od navedenih HPC sistema, zbog potrebe efikasne realizacije posla jer za uspešan i dalji razvoj sistema visokih performansi, neophodan je razvoj i drugih naučnih oblasti. Kao argument tome jeste da trenutna industrija ne može da ispuni očekivani rast broja potrebnih servera, sa postojećom tehnologijom rednog dodavanja servera. Glavni razlog za to je potrošnja električne energije i prostorni zahtev, za podrškom dodatnih računarskih centara. Realizacija novih računarskih centara zahteva upotrebu novih tehnologija, modularnog pristupa i razvoja novih tipova servera sa većim stepenom efikasnosti. Privatni HPC super računari će u budućnosti biti suočeni sa izazovom održivosti

upotrebe prostora i njihove korisnosti, sa glavnim pitanjem da li mogu da pokrivaju sopstvene troškove za funkcionisanje celokupnog sistema.

Dobijeni su određeni rezultati prilikom izgradnje laboratorijskog okruženja upotrebom HPC operativnog sistema gde smo mogli da vidimo osnovne mogućnosti na osnovu kojih možemo da planiramo konstruisanje našeg.

Sistemi visokih performansi u poslovnom svetu su samo skup produkcionih alata koji ispunjavaju određeni zadatak. Većina ozbiljnih platformi za proizvodnju video materijala je mešovitog karaktera, što znači da u svom radu koristi više različitih operativnih sistema. [49] Proizvodnja na nivou kreativnih aplikacija ili specifičnih render alata put Maya, Maxwell, Arnold, Nuke, Houdini, Cinema 4D, Premiere, FinalCutPro itd. podrazumeva podršku za pokretanje na jednom ili dva operativna sistema (na primer Windows, MacOS). HPC sistemi i virtualizacija rešavanju ovaj problem, pružajući korisnicima potrebno radno okruženje. Virtualizirana radnog okruženja (eng. VDI - Virtual desktop infrastructure) se najviše koriste u više korisničkom okruženju, gde postoji mešanje različitih aplikacija i operativnih sistema, i u budućnosti će se koristiti sve više.

Kada je vaš posao proizvodnja video materijala, morate biti u mogućnosti da realizujete projekat u najkraćem roku bez prekida u radu. Nažalost horizontalno skaliranje preko više servera na jednog lokaciji nije uvek moguće, zbog čega se vrši globalna distribucija virtuelnih servera (VDI okruženja) u različitim geografskim centrima, obezbeđujući konstantnu dostupnost i bezbednost podataka koji se razmenjuju. [50]

Virtuelne laboratorije koje se zasnivaju na HPC rešenjima se koriste za realizaciju složenih scenarija, koji uključuju kompleksne mrežne topologije i konstantnu povezanost na više mreža, vršeći usmeravanje na više različitih servera, u cilju dobijanja optimalnih performansi. Sposobnost da se sačuva kompletno stanje konfiguracije više mašina i da se napravi backup u trenutku je glavna korist virtualizacije, posebno ako neki projekat zavisi od tih virtuelnih mašina. Automatizacija i kontrola životnog ciklusa VDI mašina je korisna za sve video kompanije koje posluju i rade van svoj sedišta, pružajući im sve potrebne resurse bez obzira gde se nalaze, što znatno smanjuje troškove prilikom transporta opreme. [48]

Literatura

- [1] E. Joseph, „HPC Market Update, HPC Trends In the Oil/Gas Sector and IDC's Top 10 Predictions,“ IDC, analyze the future, 2014.
- [2] M. H. X. S. a. H. Y. Chenggang Lai, „hpcwire,“ hpcwire, 23 1 2014. [Na mreži]. Available: <http://www.hpcwire.com/2014/01/23/emerging-architectures-boost-geospatial-application-performance/>. [Poslednji pristup 02 02 2014].
- [3] I. hpc, „What is high performance computing?,“ 10 10 2014. [Na mreži]. Available: <http://insidehpc.com/hpc-basic-training/what-is-hpc/>. [Poslednji pristup 10 10 2014].
- [4] D. Lecomber, Allinea, I. Phillips, ARM, F. Subirada, BSC, F. Bodin, CAPS, J. Gonnord, CEA, S. Bassini, CINECA, G. Tecchiolli, Eurotech, G. Lonsdale, Fraunhofer, A. Pflieger, IBM, B. Andrietti, Intel, T. Lippert, FZJ, A. Bode, LRZ, H. Falter, ParTec, P. Blouet, ST i M. Muggeridge, Xyratex, „etp4hpc.eu,“ etp4hpc.eu, 2014. [Na mreži]. Available: http://www.etp4hpc.eu/wp-content/uploads/2013/06/ETP4HPC_book_singlePage.pdf. [Poslednji pristup 05 12 2014].
- [5] H. G. Hans W. Meuer, „Supercomputers – Prestige Objects or Crucial Tools for Science and Industry?,“ *Supercomputers - top500.org*, p. 21, 2014.
- [6] PhD Douglas Eadline, Making Everything Easier! High Performance Computing, Indianapolis, Indiana: Wiley Publishing, Inc., 2009.
- [7] E. C. R. v. d. G. Victor Eijkhout, Introduction to High Performance Scientific Computing, 2nd edition, Texas: Texas Advanced Computing Center, The University of Texas at Austin, 2014.
- [8] C. R. S. Kevin Down, RISC Architectures, Optimization & Benchmarks - High performance computing 2th Edition, Cambridge, Koln, Paris, Sebastopol, Tokyo: O'REILLY, 2014.
- [9] HP, „HP ProLiant Moonshot Server,“ HPC , 2014.
- [10] I. T. s. Fabric, „Intel True Scale Fabric Architecture: Enhanced HPC Architecture and Performance,“ 2007. [Na mreži]. Available: <http://www.intel.com/content/www/us/en/infiniband/true-scale-enhanced-hpc-architecture-white-paper.html>. [Poslednji pristup 10 10 2014].
- [11] „Computer World,“ 23 10 2014. [Na mreži]. Available: <http://www.computerworld.com/article/2837885/servers/hp-gives-moonshot-server-its-first-xeon-chip.html>.
- [12] T. P. Morgan, „HP Offers Exclusive Peek Inside Impending Moonshot Servers,“ 26 10 2013. [Na mreži]. Available: <http://www.enterprisetech.com/2013/11/26/hp-offers-exclusive-peek-inside-impending-moonshot-servers/>. [Poslednji pristup 2014].
- [13] Vantrix, „Video processing on HP ProLiant,“ HP, 2014.
- [14] L. Dignan, „ZDNet.com HP launches Project Moonshot,“ 08 04 2013. [Na mreži]. Available: <http://www.zdnet.com/article/hp-launches-project-moonshot-powered-with-intels-atom-at-first/>. [Poslednji pristup 2014].
- [15] L. Reiter, „How Much do 120 teraFLOPS Cost?,“ nimbix, 2014.
- [16] H. Apollo, „http://www.hp.com/hpinfo/newsroom/press_kits/2014/HPDiscover2014/Apollo8000datasheet.pdf,“ 06 2014. [Na mreži]. Available:

- http://www.hp.com/hpinfo/newsroom/press_kits/2014/HPDiscover2014/Apollo8000datasheet.pdf. [Poslednji pristup 27 04 2015].
- [17 E. Turkel, „HP/NVIDIA Solutions for HPC Compute and Visualization Performance,” 2015.]
- [18 *National Renewable Energy Lab slashes data center power costs with HP servers*, 2013.]
- [19 L. Hewlett-Packard Development Company, „HP BLc 4X QDR InfiniBand Switch Installation,” Hewlett-Packard Development Company, L.P., 2009.]
- [20 HP, 2015. [Na mreži]. Available: <http://h18004.www1.hp.com/products/blades/components/infiniband/specifications.html>. [Poslednji pristup 05 2015].]
- [21 graph500.org, „The graph 500 list, Jun - November 2014,” 06 2014. [Na mreži]. Available: http://www.graph500.org/results_jun_2014. [Poslednji pristup 04 2015].]
- [22 IBM, „IBM Blue Gene/Q,” IBM Systems Technical Computing HPC Solutions, 06 2015. [Na mreži]. Available: <http://www-03.ibm.com/systems/technicalcomputing/solutions/bluegene/index.html>. [Poslednji pristup 06 2015].]
- [23 M. Stephan, JUQUEEN: Blue Gene/Q - System Architecture, Mitglied der Helmholtz-Gemeinschaft, 2012.]
- [24 B. K. Gary Lakner, IBM System Blue Gene Solution, Blue Gene Q System Administration, IBM & International Technical Support Organization, 2013.]
- [25 I. f. D. a. R. i. I. S. Computing, "idris.fr," Turing, IBM Blue Gene/Q: Hardware configuration, Institute for Development and Resources in Intensive Scientific Computing, 2015. [Online]. Available: <http://www.idris.fr/eng/turing/hw-turing-eng.html>. [Accessed 07 2015].]
- [26 „Best Practice Guide – Blue Gene/Q,” prace-ri.eu, 2015. [Na mreži]. Available: <http://www.prace-ri.eu/best-practice-guide-blue-gene-q-html/>. [Poslednji pristup 2015].]
- [27 I. U. States, „IBM United States Prices, 212-478,” IBM, US, 2012.]
- [28 K. B. C. G. d. S. H. H.-E. K. K. P. P. P. S. S. Dino Quintero, IBM Power Systems 775 for AIX and Linux HPC, t. October 2012, USA: IBM Red Books - International Technical Support Organization, 2012.]
- [29 IBM Power 775 - Power Systems - Data Sheet, Somers, NY 10589: IBM Corporation, 2011.]
- [30 top500.org, „POWER 775, POWER7 8C 3.836GHZ, CUSTOM INTERCONNECT,” 06 2015. [Na mreži]. Available: <http://www.top500.org/system/177563>. [Poslednji pristup 06 2015].]
- [31 Cray, „Cray Corporate Fact Sheet,” Cray Inc., 2014.]
- [32 J. Burt, „Cray's New XC40 Supercomputers Powered by Latest Intel Chips,” EWeek, 2014.]
- [33 „Cray XC40™ Series Specifications,” Cray Inc., 2015.]
- [34 C. Merriman, „Cray XC40 supercomputer could challenge world's fastest title,” The Inquirer, 2014.]

-
- [35 N. Hemsoth, „Cray XC40 - Cray Strikes Balance with Next-Generation XC40 Supercomputer,“ 30 09 2014. [Na mreži]. Available: <http://www.hpcwire.com/2014/09/30/cray-strikes-balance-next-generation-xc40-supercomputer/>. [Poslednji pristup 07 2015].
- [36 Cray, „Cray® XC40-AC™ Specifications ProductBrief,“ 2015.]
- [37 P. d. A. J. Prof. dr Mladen Veinović, Računarske mreže, Beograd: Univerzitet Singidunum, Fakultet za informatiku i računarstvo, 2011.]
- [38 M. P. M. M. W. H. C. Brown, Editorial Guide Connectivity's crucial role in high-speed networking, BEI Componets for Connected Planet, 2015.]
- [39 ITU, „The State of Broadband 2014: Broadband for all, A Report by The Broadband Commission,“ The Broadband Commission, Switzerland, Geneva, 10.2014.]
- [40 I. i. R. B. M. D. v. K. i. i. R. B. Ir. Tommy van der Vorst, „Fast Forward - How the speed of Internet will develop between now and 2020,“ TU/E, Dialogic - Innovatie, interactie, Utrecht, 06.2014.]
- [41 L. L. N. L. Blaise Barney, Message Passing Interface (MPI), tom. %1 od %2UCRL-MI-133316, llnl.gov, 2014.]
- [42 L. L. N. L. Blaise Barney, Introduction to Parallel Computing, llnl.gov, 2014.]
- [43 G. E. F. G. B. T. A. J. J. D. J. M. S. V. S. P. K. B. B. A. L. R. H. C. D. J. D. R. L. G. a. T. S. W. Edgar Gabriel, „Open MPI: Goals, Concept, and Design of a Next Generation MPI Implementation.,“ *Innovative Computing Laboratory, University of Tennessee, Open System Laboratory, Indiana University, Advanced Computing Laboratory, Los Alamos National Lab*, 09 2004.]
- [44 B. T. William J. Dally, Principles and Practices of Interconnection Networks, Morgan Kaufmann, 2004.]
- [45 M. T. Inc, 2003. [Na mreži]. Available: http://www.mellanox.com/pdf/whitepapers/IB_Intro_WP_190.pdf. [Poslednji pristup 05 03 2015].]
- [46 K. B. H. K. J. K. N. M. G. N. F. P. (. S. A. V. Octavian Lascu, HPC Clusters Using InfiniBand on IBM Power Systems Servers, 10: IBM - International Technical Support Organization, 2009.]
- [47 M. Technologies, Deploying HPC Cluster with Mellanox InfiniBand Interconnect Solutions - Reference Design, tom. %1 od %2Document Number: MLNX-15-623, Sunnyvale, CA 94085: Mellanox® Technologies, 2014.]
- [48 PolyServe, High-Availability Server Clustering for E-Business, Berkeley, California 94710: PolyServe, 2015.]
- [49 D. Girard, „Enter the render queue manager,“ Arstechnica - TECHNOLOGY LAB, 12 05 2014. [Na mreži]. Available: <http://arstechnica.com/information-technology/2014/05/how-to-network-lots-of-dumb-computing-muscle-in-a-fast-efficient-render-farm/6/>. [Poslednji pristup 08 2015].]
- [50 K. Pearlman, Cutting Rhythms: Shaping the Film Edit, Karen Pearlman, 2009.]

Spisak slika

Slika 1.	Broj HPC sistema u upotrebi	12
Slika 2.	Sistemi udeo upotrebe HPC sistema	12
Slika 3.	Moonshot 1500 kućište.....	13
Slika 4.	Zadnji deo Moonshot 1500 kućišta	14
Slika 5.	HP ProLiant m710 server	16
Slika 6.	Sklopljen Apollo sistem	17
Slika 7.	HP Apollo 8000, blok za snabdevanje i konverziju električne energije..	18
Slika 8.	HP Apollo f8000 rack.....	19
Slika 9.	HP Apollo 8000 iCDU Rack	20
Slika 10.	CDU jedinicu za upravljanje protokom tople i hladne vode.	20
Slika 11.	Sistem vodenog hlađenja sa CDU jedinicom.	21
Slika 12.	Primer funkcionisanja CDU sistema, unutar jednog data centra.....	21
Slika 13.	Primer, sistem za odvođenje toplote sa jednog servera.	22
Slika 14.	HP ProLiant XL730f server.....	23
Slika 15.	HP BLc 4X QDR InfiniBand Switch Connector Bays.....	24
Slika 16.	HP Infiniband switch Administracija.	26
Slika 17.	Sistem IBM Blue Gene/Q.....	28
Slika 18.	Blue Gene/P compute card sa 4 jezgra i 8 MB cache memorije.	29
Slika 19.	Blue Gene/Q node card, Primer sastavljenog noda.	30
Slika 20.	Blue Gene/P node card kućište sa compute modulima.....	30
Slika 21.	Blue Gene/Q Desing, proces sklapanja sistema.	32
Slika 22.	IBM Power 775, kućište.	34
Slika 23.	Sastavljen Cray XC40	38
Slika 24.	Cray DataWarp I/O tehnologija.....	39
Slika 25.	Razlika u brzini prilikom upotrebe DataWarp I/O tehnologije.	39
Slika 26.	A-B komunikacija OSI model	44
Slika 27.	Projektovana prosečna količina saobraćaja za period od 2013-2020..	45
Slika 28.	Paralelna komunikacija upotrebom MPI protokola.....	47
Slika 29.	Primer Interconnect komunikacije	47
Slika 30.	Primer povezanosti rutera, Interconnect mreža	48
Slika 31.	InfiniBand konektor sa kablom.	49
Slika 32.	InfiniBand kabl i način spajanja pinova.	49
Slika 33.	InfiniBand PCI-E kartica	50
Slika 34.	Komunikacija unutar InfiniBand mreže.	51
Slika 35.	Mellanox InfiniBand Configurator	53

Slika 36.	Mellanox InfiniBand Configurator, Bandwidth (CBB) Configuration	53
Slika 37.	Fat-Tree topologija, 324 noda i 36-portni switch-evi.....	54
Slika 38.	Teoretske performanse HPC klastera	56
Slika 39.	HPC klaster.....	56
Slika 40.	DNS Balansiranje saobraćaja, data centri Beograd, Novi Sad.	58
Slika 41.	Lab. - Šema laboratorijskog okruženja.....	60
Slika 42.	Lab. - RocksCluster v6.2 (Sidewinder)	61
Slika 43.	Lab. - Izbor odgovarajućih servisa – Rolls.....	61
Slika 44.	Lab. - Izabrani odgovarajući servisi – Rolls.....	62
Slika 45.	Lab. - Osnovne informacije o Klasteru.	62
Slika 46.	Lab. - Podešavanje Javne mreže na kartici eth1	63
Slika 47.	Lab. - Podešavanje Privatne mreže na kartici eth0.....	63
Slika 48.	Lab. - Podešavanje Privatne mreže na kartici eth0.....	64
Slika 49.	Lab. - Podešavanje password-a za nalog root.....	64
Slika 50.	Lab. - Podešavanje vremenske zone, NTP server.	65
Slika 51.	Lab. - Particionisanje diska.	65
Slika 52.	Lab. – Ručno Particionisanje diska.	66
Slika 53.	Lab. – Ručno Particionisanje diska, dodavanje particije.	66
Slika 54.	Lab. - Instalacija klastera.....	67
Slika 55.	Lab. - Logovanje na sistem	67
Slika 56.	Lab. - Logovanje na sistem	68
Slika 57.	Lab. - Dodavanje novog noda.....	68
Slika 58.	Lab. - Dodavanje novog noda.....	69
Slika 59.	Lab. - Dodavanje nodova.....	69
Slika 60.	Lab. – Osnovna konfiguracija HPC sistema sa HeadNodom.....	70
Slika 61.	Lab. – Osnovna konfiguracija HPC sistema HeadNod i 6 nodova.	70
Slika 62.	Lab. – Pregled opterećenja po nodovima.	71
Slika 63.	Lab. – Pregled hardverskih mogućnosti singidunum-hpc.lokal.	71
Slika 64.	Lab. – Virtualizacija sistema.	72
Slika 65.	Lab. – Virtualizacija sistema #2.	72

Slike koje su korišćene u ovom radu su preuzete iz već navedene literature, osim gde je naznačeno drugačije.

Spisak tabela

Tabela 1.	Proizvođači i njihovi sistemi, sa sistemskim udelom.	11
Tabela 2.	Hardverske karakteristike - Moonshot 1500 kućišta	15
Tabela 3.	Hardverske karakteristike – HP ProLiant m710.....	16
Tabela 4.	Karakteristike HP Apollo f8000 rack-a	18
Tabela 5.	Hardverske karakteristike – HP ProLiant XL730f	24
Tabela 6.	Karakteristike HP Infiniband switch-a	25
Tabela 7.	Uporedni prikaz karakteristika HP servera.....	27
Tabela 8.	Blue Gene/Q JUQUEEN konfiguracije	33
Tabela 9.	Karakteristike jednog Blue Gene/Q noda.	33
Tabela 10.	Hardverske karakteristike – IBM Power 775.....	35
Tabela 11.	Uporedni prikaz karakteristika IBM servera	36
Tabela 12.	Hardverske karakteristike - Cray XC40	39
Tabela 13.	Hardverske karakteristike - Cray XC40-AC	41
Tabela 14.	Uporedni prikaz karakteristika Cray XC servera	42
Tabela 15.	Brzine prenosa InfiniBand arhitekture.	50
Tabela 16.	Balansiranje saobraćaja domena upotrebom više servera.	58
Tabela 17.	Specifikacija laboratorijskog okruženja.	60

Spisak formula

Formula 1.	Formula za izračunavanje brzine izražene u FLOPS-ima.	55
Formula 2.	Performanse po serveru za procesor Intel E5-2600.....	55