Klasteri računara i skalabilno paralelno računanje

- Arhitekura Klastera
- Midlver
- JMS

Računarski klaster

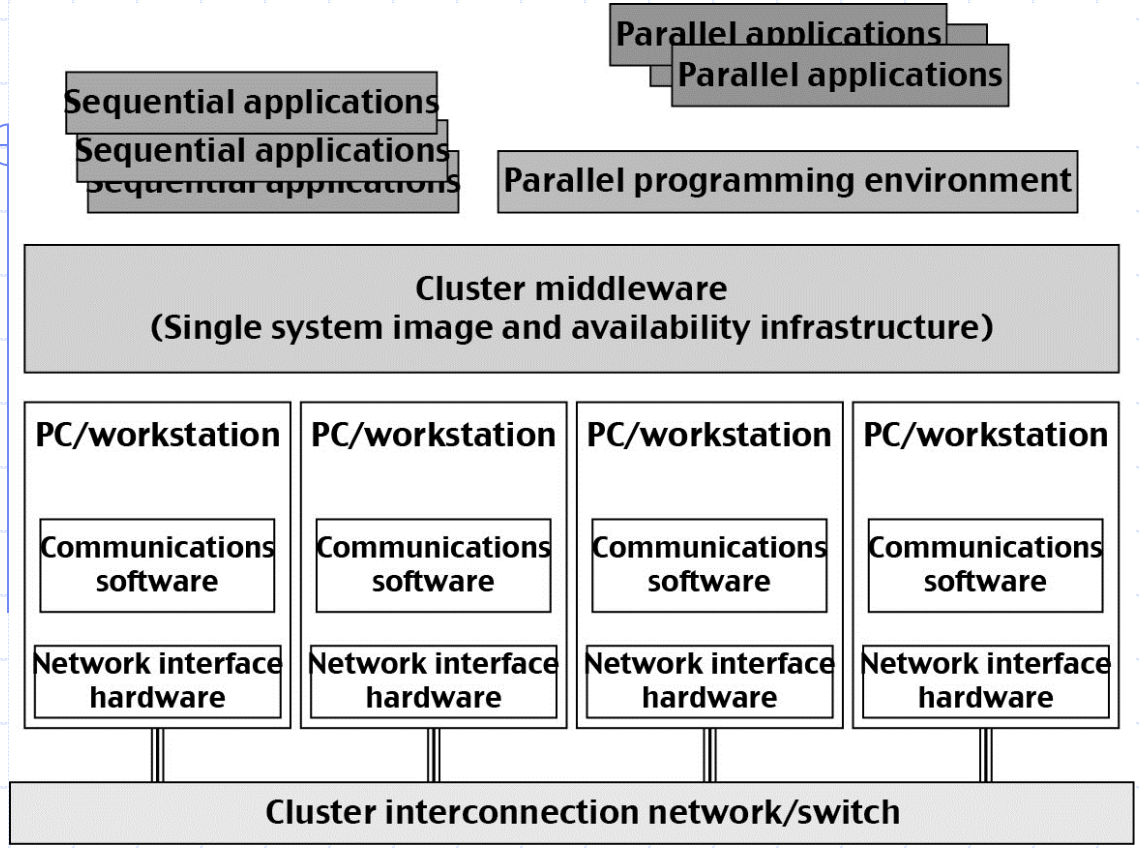
- Kolekcija međusobno povezanih računara, koji rade zajedno kao jedan, integrisan rač. resurs
  - Klaster podržava paralelizam na nivou posla i distribuirano računanje sa visokom raspoloživošću
  - Midlver na bazi slanja poruka

- Tipičan klaster:
  - Spajanje više slika u SSI (single-system image)
  - Koristi komunikacione protokole sa malim kašnjenjima
  - Sprega sa SSI slabija od sprege u SMP

Vrste klastera

- Postoje sledeće vrste klastera:
  - PC klaster (uglavnom Linux klasteri)
  - Klasteri radnih stanica (Cluster of Workstations, COW)
  - Klaster servera ili Farma servera
  - Klasteri SMP ili ccNUMA (cache coherent Non-Uniform Memory Architecture) sistema
  - Klasteri MPP procesora (MPP = massively parallel processors); 85% prvih-500 sistema su na bazi MPP

Arhitektura Klastera



- Slojevi arhitekture:
  - Aplikativni (okruženje za prog., UI, baza podataka, OLTP, itd.)
  - Midlver (SSII + Infrastruktura za visoku raspoloživost)
  - Lokalni OS i komunikacioni softver na čvorovima PC/WS
  - Mreža za povezivanje/komutaciju

Atributi za klasifikaciju klastera

Atributi	Vrednost atributa	
Pakovanje	Kompaktno	Labavo (Slack)
Upravljanje	Centralizovano	Decentralizovano
Homogenost	Homogen	Heterogen
Zaštita	Zatvoren	Izložen
Primer	Namenski klaster	Klaster preduzeća

Klasifikacija Klastera (1/2)

- Skalabilnost:
  - Mogućnost dodavanja servera u klasteru, ili dodavanje više klastera u mreži
- Pakovanje: Kompaktno / Labavo
  - Kompaktno – pakovanje u stalcima u mašinskoj sobi
  - Labavo – geografski distribuirani PC, ili radne stanice
- Upravljanje: Centralizovano / Decentralizovano

Klasifikacija Klastera (2/2)

- Homogenost: Iste naprama različite platforme (tj. CPUs, OSs)
- Programabilnost: Sposobnost da izvršava različite aplikacije
- Zaštita: Izloženost komunikacije između klastera

Kritični aspekti projektovanja klastera i izvodivost implementacije

Osobina	Funkcionalna karakterizacija	Izvodive implementacije
Raspoloživost i podrška	HW i SW podrška za održivu visoku raspoloživost (HA)	Preuzimanje, kontrolna tačka, opravak vraćanjem nazad, nonstop OS, itd.
HW otpornost na otkaze	Automatizovano rukovanje otkazima i otklanjanje svih kritič. tačaka otkaza	Redundantnost komponenti, vruće razmene, RAID, više izv.napajanja, itd.
SSI (Single System Image)	SSI na funkcionalnom nivou HW, SW, midlvera i OS proširenja	HW ili midlver mehanizmi za postizanje DSM sa koherentnim keš mem.
Efikasna komunikacija	Smanjenje režije slanja poruka i skrivanje kašnjenja	Brzo slanje poruka, aktivne poruke, poboljšana MPI biblioteka, itd.
Rukovanje poslovima na celom klasteru	Globalni sistem za rukovanje posl. sa boljim raspoređivanjem i nadzorom	Primena sistema za ruk. poslovima kao što je LSF, Codine, itd.
Dinamičko uravnote-ženje opterećenja	Uravnoteženje opterećenja svih čvorova sa oporavakom od otkaza	Nadzor opterećenja, migracija procesa, replikacija poslova, grupno raspored.
Skalabilnost i programabilnost	Dodavanje servera i klastera sa pove. opterećenja i skupova podataka	Skalabilna međuveza, nadzor perform., distribuirano izvršenje i bolji SW alati

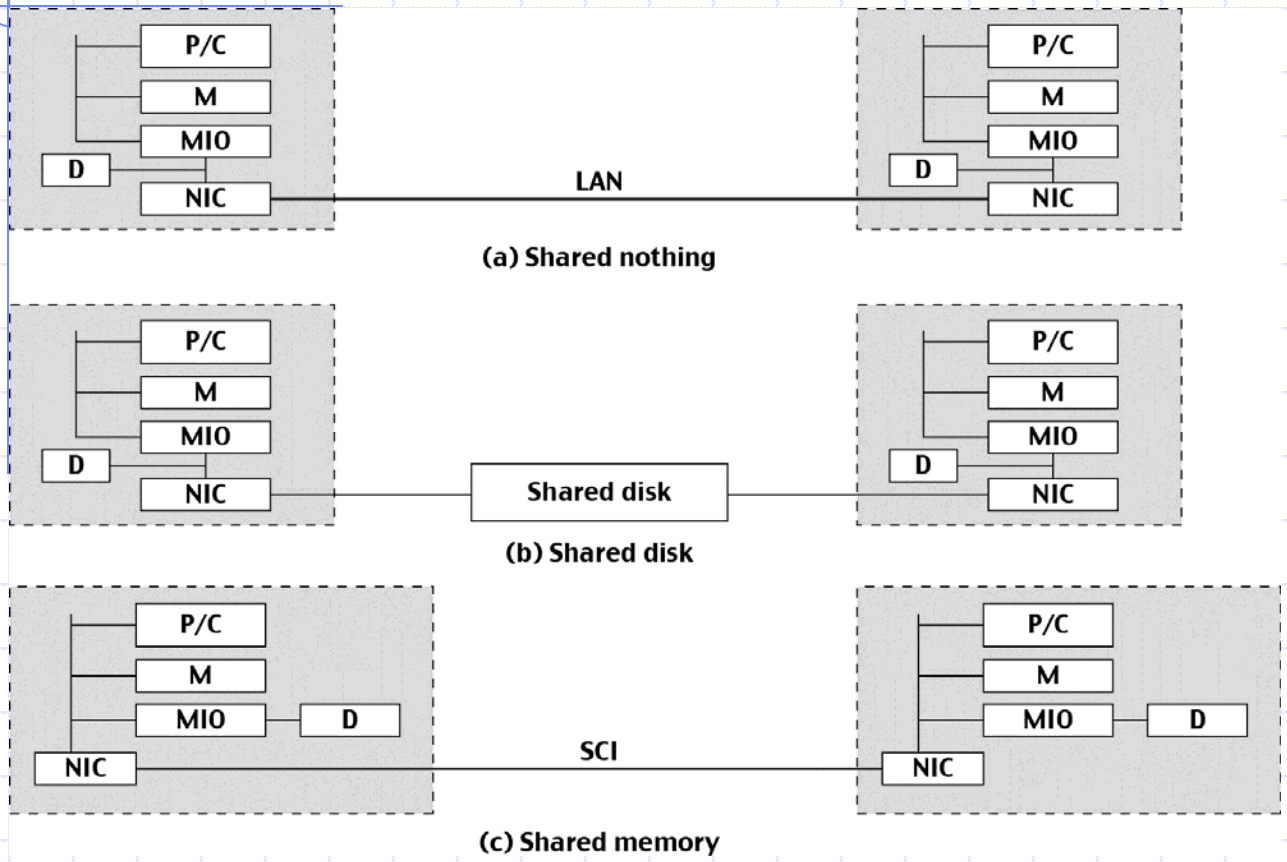
Prednosti klasterizacije

- Visoka raspoloživost (HA)
- Hardverska otpornost na otkaze
- Pouzdanost OS i aplikacije
  - Izvršava se više kopija OS i aplikacija
- Skalabilnost
- Visoka performansa
  - Veći propusnost (throughput) sistema

Prvih 5 superračunara u 2010

Ime sistema	Opis arhitekture (broj jezgara, procesor, GHz, OS, topologija međuveze)	Stalna brzina	Snaga po sistemu
Jaguar u Oak Ridge Nac. Lab u SAD	Cray XT5-HE: je MPP sa 224,162 jezgra na 2.6 GHz Opteron 6-jezgarni procesori, međuveza je 3-D torus	1.759 PFlops	6.95 MW
Nebulae u kineskom Nac. Centru (NSCS)	Dawning TC3600 Blade sistem: sa 120,640 jezgara u 2.66 GHz Intel EM64T Xeon X5650 i Nvidia GPU sa Linux, međuveza je Infiniband QDR mreža	1.271 PFlops	2.55 MW
Roadrunner u DOE/NNSA/LANL SAD	IBM BladeCenter QS22/LS21 klaster od 122,400 jezg. u 12,960 3.2 GHz POWER XCell 8i i 6,480 AMD 1.8 GHz Opteron procesora, sa Linux i Infiniband mrežom	1.042 PFlops	2.35 MW
Kraken XT5 u NICS Univerzitet Tenesi	Cray XT5-HE: je MPP sa 98,928 jezgra na 2.6 GHz Opteron 6-jezgarni procesori, međuveza je 3-D torus	831.7 TFlops	3.09 MW
JUGENE u FZJ, Nemačka	IBM BlueGene/P sa 294,912 procesora: PowerPC jezg. 4-way SMP čvorovi i 144 TB memorije u 72 stalka, međuveza je 3-D torus	825.5 TFlops	2.27 MW

Tri načina povezivanja čvorova klastera



- Tri načina:
- (a) bez deljenih resursa
- (b) sa deljenim diskom
- (c) sa deljenom memorijom

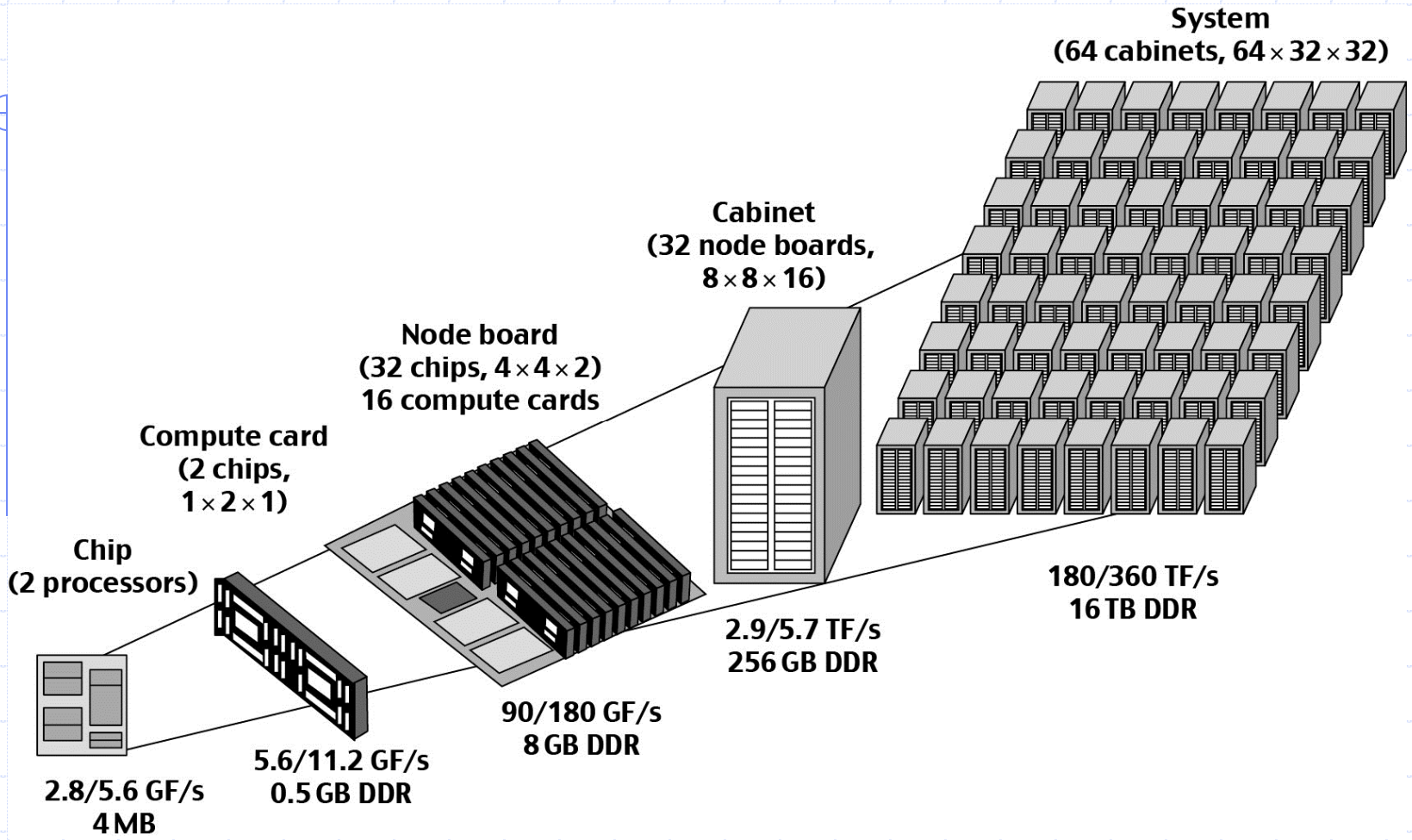
Važnije arhitekture računskih čvorova za klastere iz 2010

Arhitektiura čvora	Važnije karakteristike	Reprezentativni sistemi
Homogeni čvor sa istim višejezgarnim procesorima	Više višejezgarnih procesora koji su povezani preko ukrsne matrice (crossbar) sa deljenom mem. i lokalnim diskovima	Cray XT-5 koristi 2 6-jezgarna AMD Opteron procesora u svakom čvoru
Hibridni čvorovi sa CPU plus GPU ili FLP za ubrzanje	Opštenamenski CPU za celobrojne operacije + GPU kao koprocesor za operacije u pokretnom zarezu (FLP)	Kineski Tianhe sistem koristi 2 Intel Xeon procesora + 2 AMD GPU u svakom čvoru

Primer: Superračunar IBM Blue Gene/L

- Razvili IBM i Lawrence Livermore National Lab
- Na privih 17 od 100 pozicija u rangiranju na top500.org, uključujući 5 u prvih 10
  - Teorijska vršna brzina: 360 TeraFLOPS
- Najveća konfiguracija:
  - U Lawrence Livermore Nat'l Lab
  - 64 fizičkih stalaka
  - 65,536 računskih čvorova
  - Međuračunarska mreža tipa Torus od 64 x 32 x 32

Arhitektura IBM BlueGene/L HW-a



- Najbrži MPP u svetu u 2005
- Napravili IBM i LLNL timovi i finansirano od US DoE ASCI

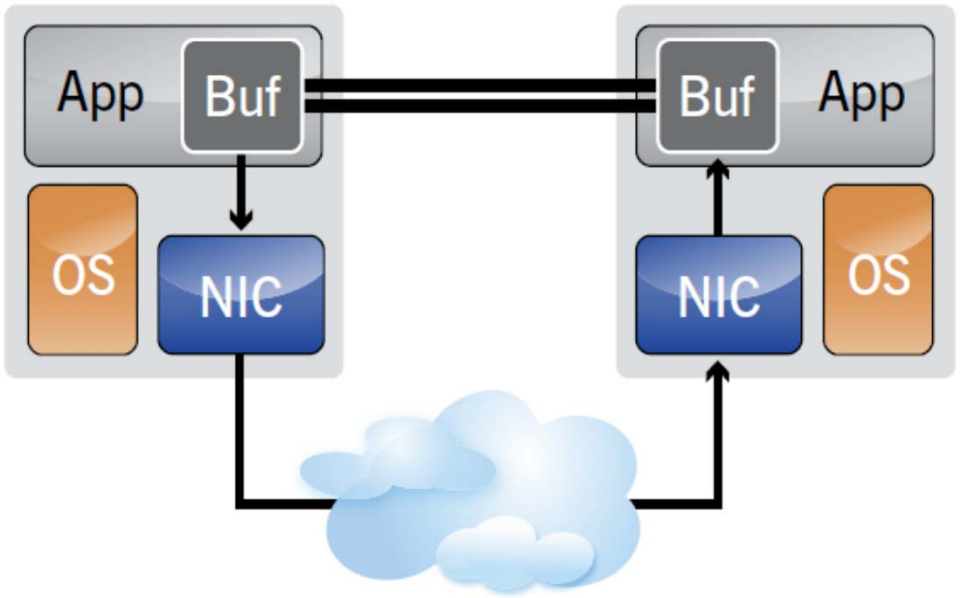
Poređenje 4 tehnologije međuveza za klaster

Osobina	Myrinet	Quadrics	InfiniBand	Ethernet
Brzine linkova	1.28 Gb/s (M-XP) 10 Gb/s (M-10G)	2.8 Gb/s (QsNet) 7.2 Gb/s (QsNetII)	2.5 Gb/s (1X) 10 Gb/s (4X) 30 Gb/s (12X)	1 Gb/s
MPI kašnjenje	~3 us	~3 us	~4.5 us	~40 us
Mrežni procesor	Da	Da	Da	Ne
RDMA	Da	Da	Da	Ne
Topologije	Bilo koja	Bilo koja	Bilo koja	Bilo koja
Unutrašnja topologija	Clos	Plitko stablo	Plitko stablo	Bilo koja
Usmeravanje	Na bazi izvora	Na bazi izvora	Na bazi odredišta	Na bazi odredišta
Kontrola toka	Stani-kreni (Stop and Go)	Mravlje putanje (Worm-hole)	Apsolutni kredit (Absolute credit)	802.3x

Primer: InfiniBand (1/2)

- Obezbeđuje jednostavnu uslugu za rukovanje porukama (messaging)
  - Direktan pristup usluzi za rukovanje porukama bez potrebe za oslanjaje na OS
- Usluga stvara kanale koji povezuju aplikacije
  - Ovi kanali su između virtuelnih adresnih prostora aplikacija

Primer: InfiniBand (2/2)

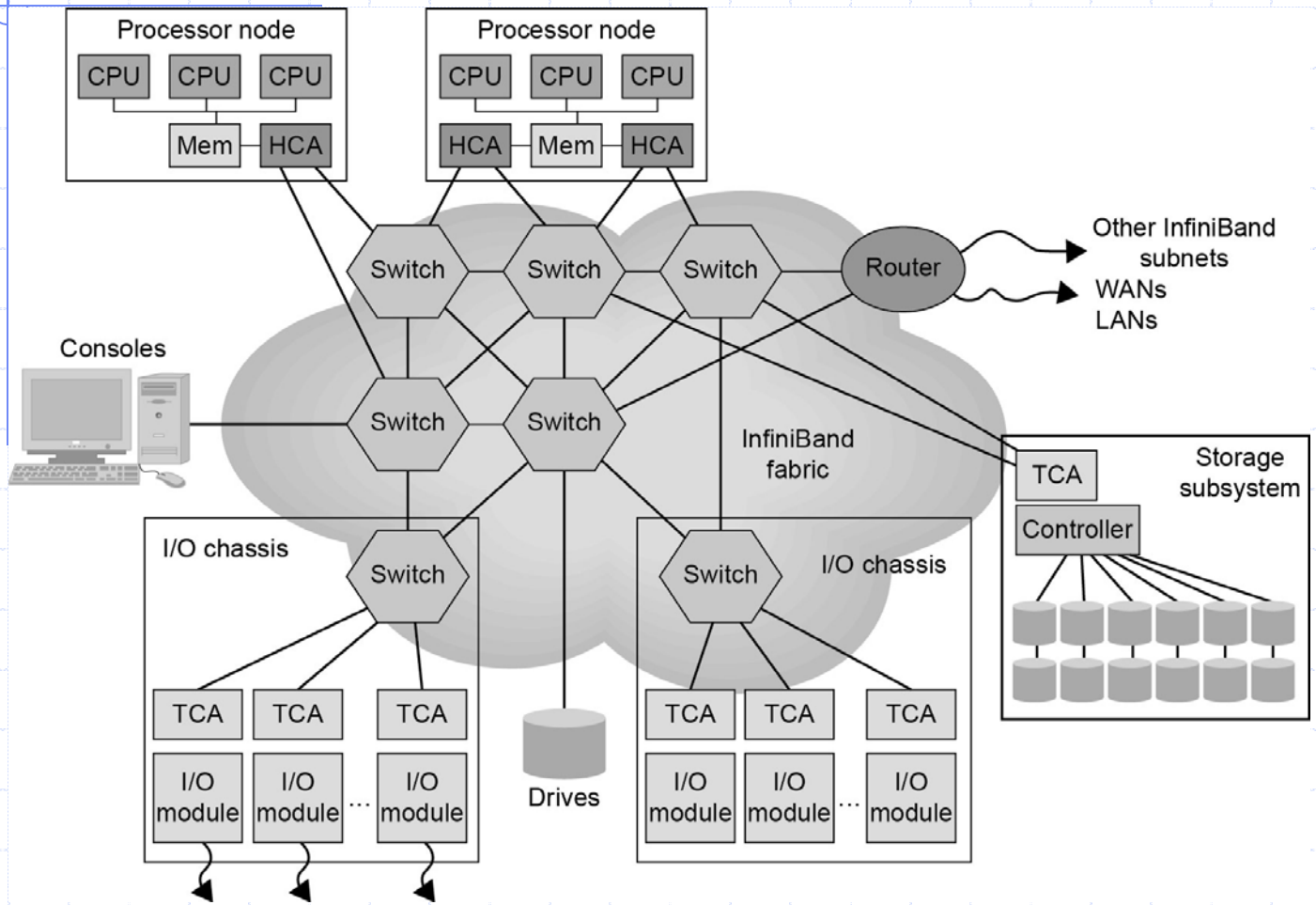


- InfiniBand stvara kanal koji direktno povezuje virtuelne adresne prostore aplikacija
- Te dve aplikacije mogu biti u razdvojenim fizičkim adresnim prostorima – u različitim serverima

InfiniBand arhitektura (1/2)

- HCA (Host Channel Adapter)
  - Povezuje krajnji čvor, kao što je server ili skladišni uređaj, na InfiniBand mrežu
- TCA (Target Channel Adapter)
  - Poseban HCA namenjena za upotrebu u ugrađenom okruženju kao što je skladišni uređaj
- Komutator (switch)
  - Projektovan da dostigne ciljnu performansu i troškove
- Usmerivač (router)
  - Za segmentaciju veoma velike mreže u manje podmreže povezane usmerivačima

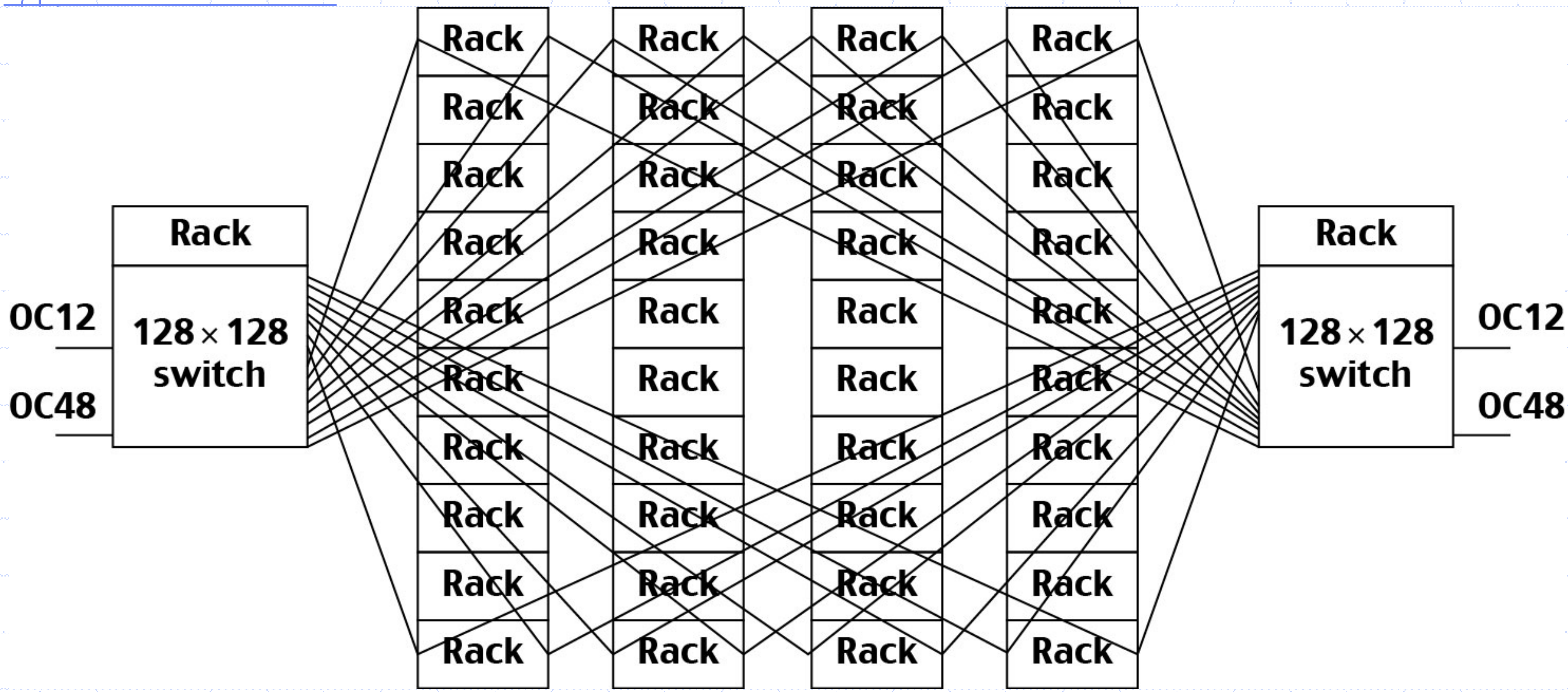
InfiniBand arhitektura (2/2)



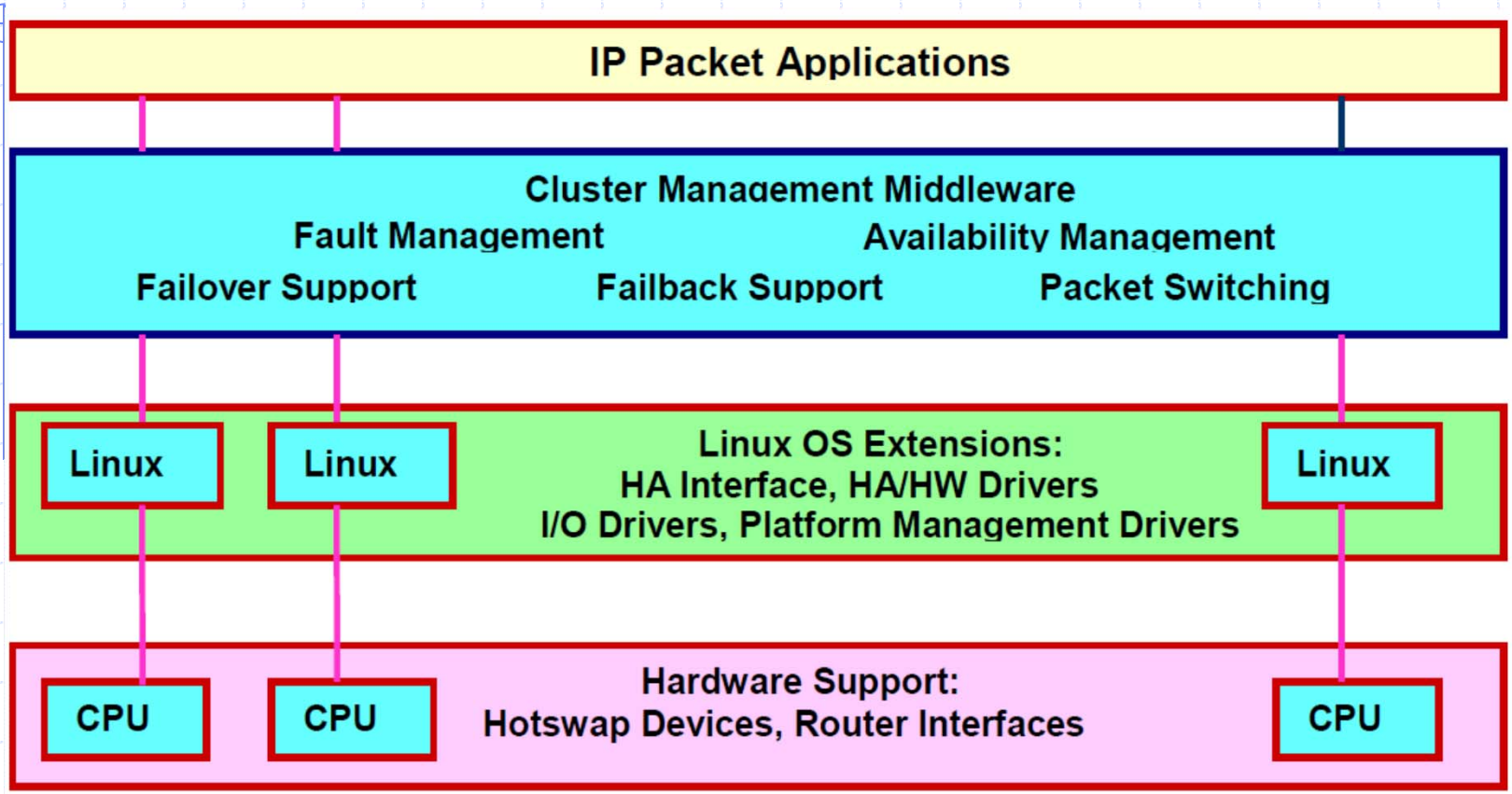
Primer: Google Search Engine

- Superklaster nad brzim PC rač. i Gigabit LAN za aplikacije globalnog pretraživanja web stranica
  - Klaster je smešten u 40 PC/switch stalaka sa 80 PC po stalku i 3200 PC ukupno
  - U dva stalka je smešteno dva 128 x 128 Gigabit Ethernet komutat., prednji (front) rač., UPS, itd.
- Google SW sistemi za paralelne pretrage, URL povezivanje, rangiranje strana, rukovanjem datotekama i bazama podataka, itd.

Arhitektura klastera Google Search Engine



Slojevi arhitekture klastera: Hardver, Softver, i Midlver



- Midlver (rukovanje otkazima i raspoloživošću), Linux proširenja, i hardver za postizanje visoke raspoloživosti u klasterskom sistemu

Projektantski principi za klastere

- Glavni projektantski principi za klastere su:
  - Jedna sistemska slika (SSI)
  - Visoka raspoloživost (HA)
  - Otpornost na greške (FT)
  - Oporavak povratkom unazad (Rollback recovery)

Jedna sistemska slika (SSI)

- Jedna sistemska slika je iluzija, stvorena pomoću SW ili HW, koja predstavlja kolekciju resursa kao jedan integrisani moćni resurs
  - SSI čini da se klaster pojavljuje korisniku kao jedna mašina, sa aplikacijama, koja je povezana na mrežu
  - Klaster sa više sistemskih slika nije ništa drugo do zbirka nezavisnih računara (tj. opšti distribuirani sistem)

Osobine SSI (Single-System-Image)

- Jedan sistem:
  - Korisnici vide ceo klaster kao jedan sistem
- Simetrija:
  - Usluge su dostupne sa bilo kog čvora, tj. simetrične su ka svim čvorovima i svim korisnicima
- Lokacijska transparentnost:
  - Korisniku nije poznata lokacija uređaja koji obezbeđuje uslugu

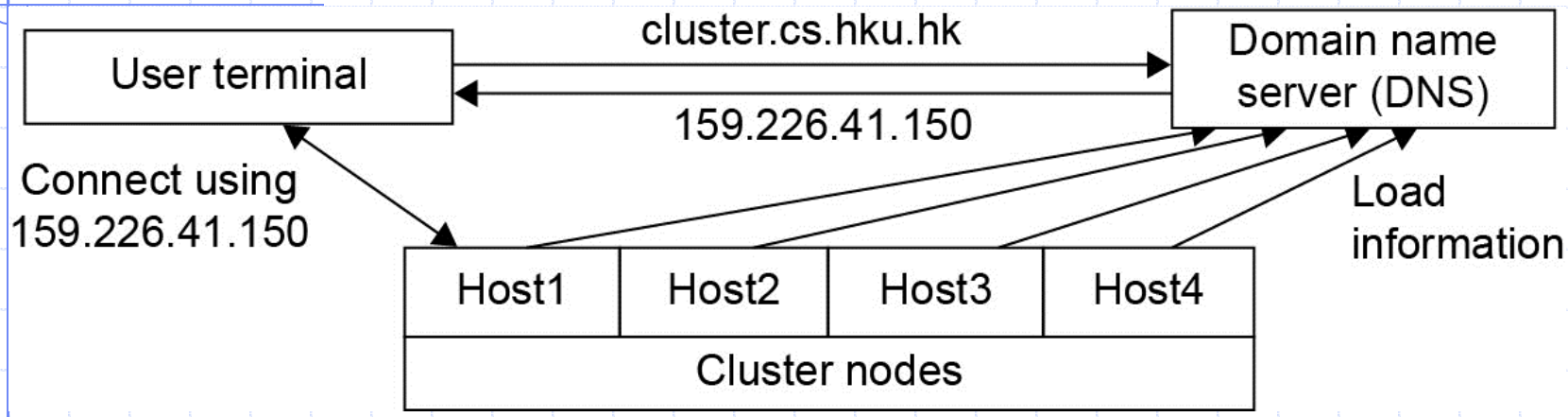
SSI usluge

- Osnove SSI usluge:
  - Jedna ulazna tačka, npr. telnet cluster.usc.edu
  - Jedna hijerarhija datoteka: xFS, AFS, Solaris MC Proxy
  - Jedan U-I, pristup mreži, i memorijski prostor
- Dodatne SSI usluge:
  - Jedno rukovanje poslovima: GlUnix, Codine, LSF, itd.
  - Jedna korisnička sprega: Kao CDE u Solaris/NT
  - Jedan prostor procesa

Projektantski zahtevi za SSI

- Zahtevi za SSI:
  - Jedna tačku upravljanja
  - Jedan adresni prostor
  - Jedan sistem za upravljanje poslovima
  - Jedna korisnička sprega
  - Jedno upravljanje procesima

Primer: Realizacija jedne ulazne tačke u klasteru računara



- 4 čvora u klasteru su domaćini za prijem korisnič. zahteva za login
- Login na klaster sa standardnom Unix komandom kao što je "telnet cluster.cs.hku.hk", gde je domen simboličko ime klastera
- DNS prevodi simboličko ime, i vraća IP adresu 159.226.41.150 najmanje opterećenog čvora, koji je u ovom slučaju čvor Host1
- Korisnik se onda prijavljuje koristeći ovu IP adresu
- DNS periodično prima informaciju o opterećenju od čvorova domaćina u cilju uravnoteženja opterećenja (load-balancing)

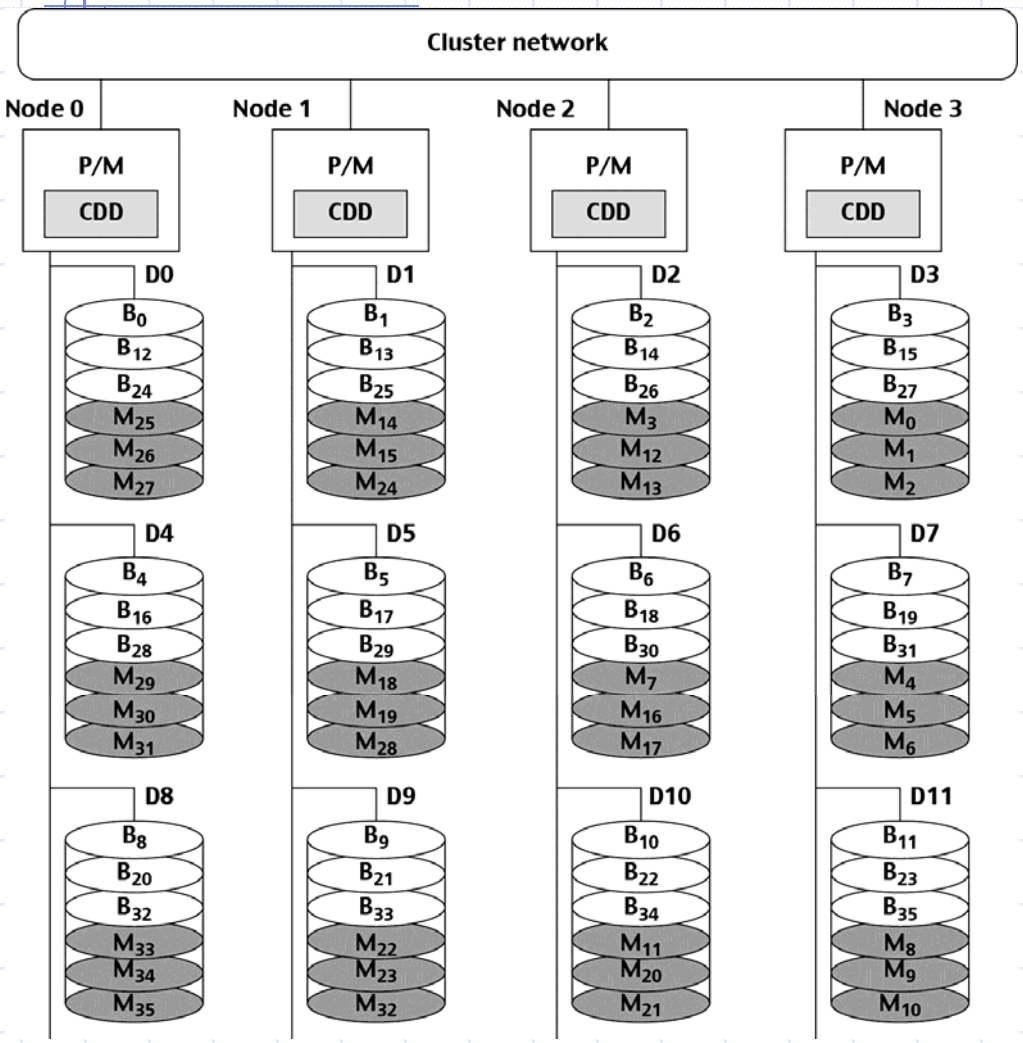
Jedna hijerarhija datoteka

- Iluzija jedne slike sistema dat. koja integriše lokalne i globalne diskove i druge uređaje sa datotekama
- Datoteke na 3 tipa lokacija u klasteru:
  - Lokalno skladište - disk na lokalnom čvoru
  - Udaljeno skladište – diskovi na udaljenim čvorovima
  - Stabilna skladišta - osobine:
    - Perzistentnost - podaci ostaju u njemu neki period vremena (npr., 7 dana) i nakon potpunog isključenja klastera
    - Otpornost na greške – korišćenjem redundantnih uređaja i periodičnim bekapima na trake

Stabilno skladište

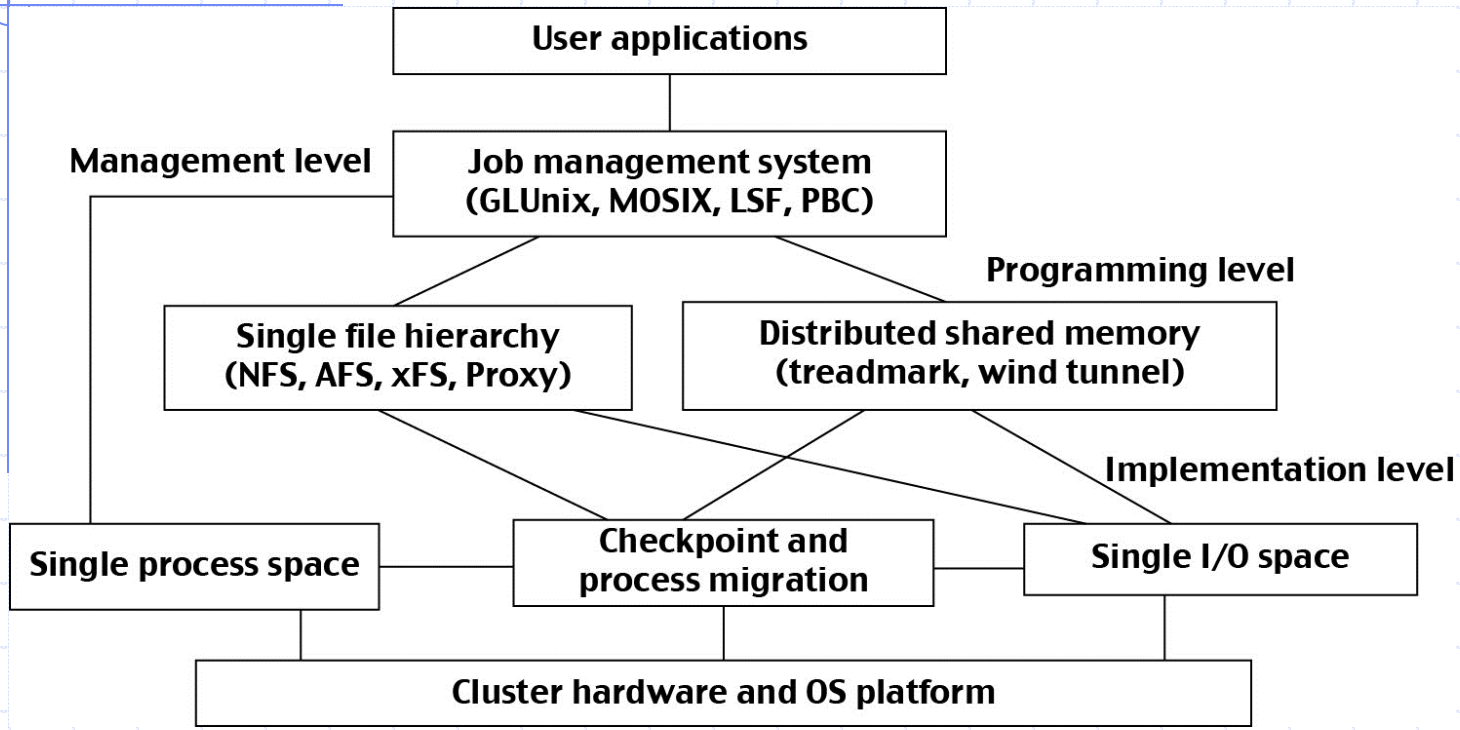
- Može se realizovati:
  - Centralizovano (jedan RAID disk)
  - Distribuirano (lokalnih diskova u čvorovima klastera)
- Prednosti i nedostaci:
  - Prvi pristup sa jednim diskom ima jednu tačku otkaza i potencijalno usko grlo za performansu
  - Drugi pristup je teže realizovati, ali on može biti ekonomičniji, efikasniji, i sa većom raspoloživošću

Distribuirana RAID arhitektura



- RAID sa jednim U-I prostorom preko 12 distribuiranih diskova na 4 čvora u Linux klasteru
- Di označava disk i
- Bj označava disk blok j
- Mj je slika (tamne ploče) od bloka Bj
- P/M se odnosi na čvor procesor / memorija
- CDD (cooperative disk driver) je kooperativni disk drajver

Arhitektura midlvera za SSI klasterizaciju

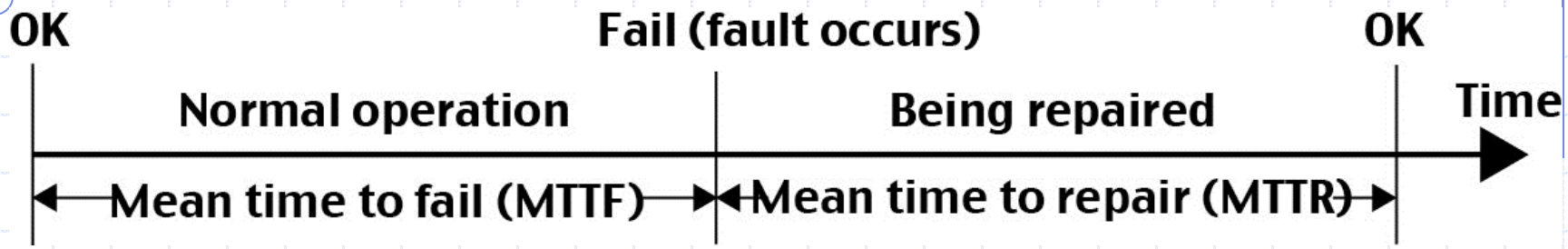


- Tri nivoa midlvera:
  - Nivo rukovanja poslovima
  - Nivo programiranja
  - Nivo implementacije

Visoka raspoloživost kroz redundantnost

- Tri pojma često idu zajedno: pouzdanost, raspoloživost, i servisibilnost (RAS)
- Definicije ovih pojmova su:
  - Pouzdanost je mera koliko dugo sistem može da radi bez kvara
  - Raspoloživost se iskazuje kao procenat vremena u kom je sistem raspoloživ za korisnika
  - Servisibilnost se odnosi na jednostavnost servisiranja sistema (održavanje, popravke, nadogradnje, itd.)

Raspoloživost i stopa otkaza



- Slika prikazuje ciklus rad-popravka (rač.) sistema
- Raspoloživost = MTTF / (MTTF + MTTR)

- Skorašnje istraživanje o Fortune 1000 kompanijama:
  - Srednji br. otkaza računara je 9 puta godišnje sa srednjim vremenom van rada od 4 sata
  - Srednji gubitak prihoda po satu van rada je $82,500

Raspoloživost nekih tipova računarskih sistema

Tip sistema	Raspoloživost (%)	Van rada u toku 1 god
Konvencionalna radna stanica	99	3.6 dana
HA sistem (high-availability)	99.9	8.5 sati
Sistem elastičan za greške (fault-resillient)	99.99	1 sat
Sistem otporan na greške (fault-tolerant)	99.999	5 min

Konfiguracije klastera otporne na otkaze

- Sledeće 3 konfiguracije se često koriste:
  - Vruća rezerva (Hot Standby)
    - Primarna komponenta obezbeđuje uslugu, dok redundantna rezervna komponenta čeka bezposlena, ali je pripravna (hot)
  - Međusobno preuzimanje
    - sve komponente aktivno rade koristan posao; kad jedna otkaže, njen teret se preraspodeljuje na druge komponente
  - Otpornost na otkaze
    - Najskuplja konfiguracija, jer N komponenti daju performansu samo jedne komponente, za N puta veću cenu. Maskira se otkaz N–1 komponenti

Preuzimanje (Failover)

- Najvažnija osobina za komercijalne aplikacije
- Kad otkaže komponenta, ostatak sistema preuzma posluživanje od komponente u otkazu
  - Mehanizam preuzimanja obezbeđuje sl. funkcije: dijagnostiku, obaveštenje, i oporavak od otkaza
- Dijagnostika otkaza (detakcija i lociranje)
  - Tehnika „javljanje" (heartbeat), gde čvorovi klastera šalju niz specijalnih (heartbeat) poruka jedan drugom
  - Ako sistem ne prima ove poruke od nekog čvora, otkazao je ili taj čvor ili mrežna veza do tog čvora

Šeme oporavka

- Dva tipa tehnika oporavka:
  - Oporavak u nazad
    - Procesi periodično sačuvavaju konzistentno stanje (kontrolna tačka, eng. checkpoint) u stabilno skladište
    - Posle otkaza, sistem izoluje komp. u otkazu i vraća se na zadnju kontrolnu tačku, tzv. „vraćanje unazad" (rollback)
    - Lako se implementira i široko se koristi
    - Implicira uzaludno izvršenje
  - Oporavak u napred
    - Sistem koristi rezultate dijagnostike za rekonstrukciju validnog stanja sistema i nastavak izvršenja
    - Oporavak u napred zavisi od aplikacije i može zahtevati dodatni HW

Tehnike kontrolne tačke i oporavka

- Nivoi kontrolna tačke:
  - Jezgro OS, biblioteka, i aplikacija
- Moguće optimizacije:
  - Režija kontrolne tačke
  - Izbor optimalnog intervala kontrolne tačke
- Inkrementalna kontrolna tačka
  - Čuva se samo deo stanja koji se promenio od predhodne tačke
- Korisnički definisana kontrolna tačka
  - Kaže sistemu kada i šta da sačuva, a šta da ne sačuva

Raspoređivanje i rukovanje poslovima u klasteru

- Delovi sistema za rukovanje poslovima (JMS):
  - Korisnički server: korisnik može da podnosi poslove u jednom ili više redova, specificira zahteve za resursima, briše posao, i ispituje stanje posla ili reda
  - Raspoređivač poslova: raspoređuje i ulančava poslove u skladu sa tipovima poslova, zahtevima za resurse, raspoloživošću resursa, i politikama raspoređivanja
  - Rukovalac resursima: dodeljuje i prati resurse, sprovodi politike raspoređivanja, i prikuplja informaciju za obračun

JMS administracija

- Zahtevi:
  - JMS treba da može da dinamički rekonfiguriše klaster sa min uticajem na tekuće poslove
  - Prolog i epilog skripte se izvršavaju pre i posle svakog posla radi provera zaštite, obračuna, i čišćenja
  - Korisnici treba da imaju mogućnost čistog uništavanja svojih poslova
  - Administrator JMS treba da ima mogućnost čistog suspendovanja ili uništavanja bilo kog posla
    - Čistog znači da kad je posao suspendovan ili uništen, svi njegovi procesi moraju biti uključeni

Tipovi poslova u klasteru

- Tipovi poslova:
  - Serijski poslovi se izvršavaju na jednom čvoru
  - Paralelni poslovi koriste više čvorova
  - Interaktivni poslovi su oni koji zahtevaju brz odziv, i njihov ulaz-izlaz je usmeren na terminal
  - Paketski poslovi obično traže više resursa, kao što je veliki memorijski prostor i dugo CPU vreme
    - Oni se podnose u red poslova da bi bili raspoređeni kad resursi postanu raspoloživi (npr., van radnog vremena)

Karakteristike radnih opterećenja

- Važi pravilo 2:1, koje kaže da mreža od 64 radne stanice, sa dobrim JMS SW, može da, pored originalnog sekvencijalnog opterećenja, izdrži dodatno paralelno opterećenje od 32 čvora
  - Drugim rečima, klasterizacija daje superračunar u kom se polovina veličine klastera dobija besplatno!

Šeme raspoređivanja više poslova

- Poslovi mogu biti raspoređeni:
  - za rad u zadato vreme ili kad se desi zadati događaj
- Prioritet poslova se određuje na osnovu:
  - vremena podnošenja, resursa čvora, vremena izvršenja, memorije, prostora na disku, tipa posla, i identiteta korisnika
- Statički prioriteti
  - Prvi-došao, prvi-poslužen (slično sa FIFO)
  - Prioritet prema identitetu korisnika
- Dinamičkih prioriteti

Aspekti i šeme raspoređivanja

Aspekt	Šema	Ključni problemi
Prioritet posla	Bez istiskivanja	Kašnjenje visoko-prioritetnih poslova
	Sa istiskivanjem	Režija, implementacija
Zahtevanje resursa	Statički	Neuravnoteženo opterećenje
	Dinamički	Režija, implementacija
Deljenje resursa	Posvećeno	Loše iskorišćenje
	Deljenje prostora	Popločavanje, veliki posao
Raspoređivanje	Deljenje vremena	Upravljanje poslovima na bazi procesa sa režijom za smenu konteksta
	Nezavisno	Ozbiljno usporenje
	Grupno	Problematična implementacija
Nadmetanje sa stranim poslovima	Ostanak	Usporenje lokalnih poslova
	Migracija	Migracioni prag, migraciona režija

Režimi raspoređivanja (1/2)

- Posvećen režim:
  - Radi jedan posao na klasteru, i najviše jedan proces posla je dodeljen nekom čvoru u nekom trenutku
- Deljenje prostora:
  - Više poslova radi simultano na razdvojenim particijama (grupama) čvorova
  - Najviše jedan proces je dodeljen čvoru
  - Particija čvorova je dodeljena poslu, a međuveza i U-I podsistem mogu biti deljeni od strane svih poslova

Režimi raspoređivanja (2/2)

- Deljenje vremena:
  - Više korisničkih procesa se dodeljuje istom čvoru
  - Postoje sledeće politike paralelnog raspoređivanja:
    - Nezavisno raspoređivanje: Koristi OS svakog čvora za raspoređivanje procesa kao na običnim radnim stanicama
    - Grupno raspoređivanje (Gang): Raspoređuje sve procese paralelnog posla zajedno
    - Nadmetanje sa stranim poslovima: Raspoređuje istovremeno poslove klastera i lokalne poslove; lokalni poslovi bi trebali da imaju prioritet u odnosu na poslove klastera
  - Bavljenje sa lokacijom posla: Ostani ili migriraj

Aspekti šeme migracije

- Raspoloživost čvora:
  - Berkeley studija: Čak tokom vršnih sati, 60% radnih stanica u klasteru je raspoloživo
- Režija migracije:
  - Berkeley studija: usporenje čak do 2.4 puta
  - Usporenje je manje ako paralelni posao radi na klasteru dvostruke veličine
- Regrutacioni prag:
  - je količina vremena u kom čvor ostaje nekorišćen pre nego ga klaster proglasi za besposlen čvor

Osobine JMS (1/2)

- Svi podržavaju:
  - paralelne i paketske poslove
  - uravnoteženje opterećenja
  - dinamičko suspendovanje i nastavljanje korisničkih poslova
  - dinamičko dodavanje ili brisanje resursa (npr., čvorova)
- Većina podržava:
  - spregu komandne linije i grafičku korisničku spregu
  - heterogne Linux klastere

Osobine JMS (2/2)

- Neki podržavaju kontrolne tačke
- Većina ne podržava proces dinamičke migracije
  - Alternativa – statička migracija: može migrirati samo jednom i to nakon što je napravljen