



KLASTER RAČUNARA

Infrastruktura za elektronsko poslovanje

dr Miloš CVETANOVIĆ
dr Zaharije RADIVOJEVIĆ



Osobina	Server	Klaster	Superračunara	Datacentar	Skladište računara
Dizajn	- cena/performance - energatska efikasnost - redundansa - maksimalne performance	- cena - redundansa	- maksimalne performance - brza lokalna mreža - veliki broj multiprocesora	- konsolidacija - heterogenost - operativni troškovi	- cena/performance - energatska efikasnost - redundansa - operativni troškovi
Mreža	- Internet	- Internet - lokal	- lokal	- Internet - lokal	- Internet - lokal
Opterećenje	- interaktivno - paketno	- interaktivno	- paketno	- interaktivno - paketno	- interaktivno - paketno
Paralelizam	- na nivou niti	- na nivou zahteva	- na nivou niti	- na nivou zahteva - na nivou niti	- na nivou zahteva
Problemi	- specifičnost delova	- nepouzdanost delova	- specifičnost delova	- specifičnost delova	- nepouzdanost delova
Iskorišćenje	- veliko	- malo	- veliko	- veliko	- malo

MTTF servera 200 000 sati (~22 godine)

Klaster → 50 000 server → 6 servera dnevno

Otkaz diska 2-10% godišnje

Klaster → 50 000 server sa po 4 diska svaki → (pri 4% godišnje) 22 diska dnevno

Nepouzdanost servera → ponavljanje nezavršenih poslova

Nepouzdanost diskova → replikacija kompletnih podataka (ne treba RAID)

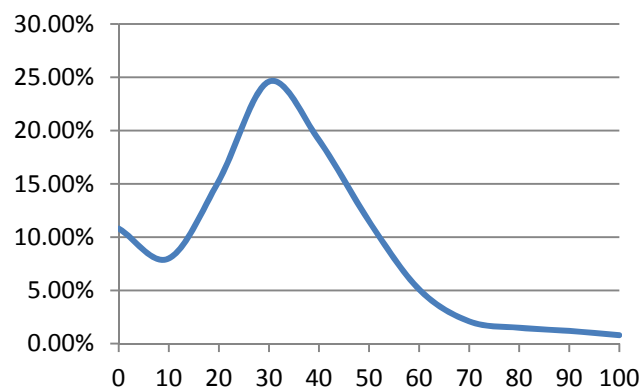
(GFS – Google File System) → Bigtable

reklasirana konzistentnost podataka (ne treba ACID)

Automatizovano nadgledanje → 1000 servera, 1 operater



Primer 8. Eksperimentalno utvrđen procenat vremena za određeni nivo iskorišćenja pojedinog servera u skladištu računara dat je grafikon kao na slici. U tabeli su dati rezultati SPECPower testova servera koji je u upotrebi u skladištu računara. Kako se menja metrika ssj_ops/watt ako se posmatra eksperimentalno utvrđeno opterećenje?



Opterećenje	Performanse	Potrošnja	Težine SPEC	Težinske performanse	Težinska potrošnja
100%	2 889 020	662	9,09%	262 638	60
90%	2 611 130	617	9,09%	237 375	56
80%	2 319 900	576	9,09%	210 900	52
70%	2 031 260	533	9,09%	184 660	48
60%	1 740 980	490	9,09%	158 271	45
50%	1 448 810	451	9,09%	131 710	41
40%	1 159 760	416	9,09%	105,433	38
30%	869 077	382	9,09%	79 007	35
20%	581 126	351	9,09%	52 830	32
10%	290 762	308	9,09%	26 433	28
0%	0	181	9,09%	0	16
Ukupno	15 941 825	4967		1 449 257	451
				ssj_ops/watt	3213

Odgovor: metrika performansi se smanjila za 24%, odnosno sa 3213 na 2445 ssj_ops/watt

Opterećenje	Performanse	Potrošnja	Težine SPEC	Težinske performanse	Težinska potrošnja	Težine iskorišćenja	Težinske performanse	Težinska potrošnja
100%	2 889 020	662	9,09%	262 638	60	0,80%	22 206	5
90%	2 611 130	617	9,09%	237 375	56	1,20%	31 756	8
80%	2 319 900	576	9,09%	210 900	52	1,50%	35 889	9
70%	2 031 260	533	9,09%	184 660	48	2,10%	42 491	11
60%	1 740 980	490	9,09%	158 271	45	5,10%	88 082	25
50%	1 448 810	451	9,09%	131 710	41	11,50%	166 335	52
40%	1 159 760	416	9,09%	105,433	38	19,10%	221 165	79
30%	869 077	382	9,09%	79 007	35	24,60%	213 929	94
20%	581 126	351	9,09%	52 830	32	15,30%	88 769	54
10%	290 762	308	9,09%	26 433	28	8,00%	23 198	25
0%	0	181	9,09%	0	16	10,80%	0	20
Ukupno	15 941 825	4967		1 449 257	451		933 820	382
				ssj_ops/watt	3213		ssj_ops/watt	2445

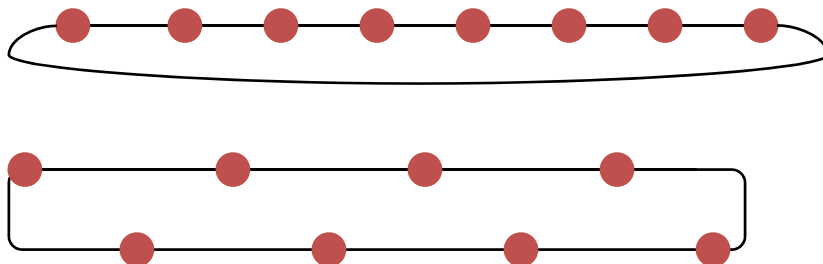


Linerani niz



Dijametar = $n-1$; prosečna udaljenost $\sim n/3$
Bisekcioni protok = 1 (u jedinicama linkova)

Torus ili prsten



Dijametar = $n/2$; prosečna udaljenost $\sim n/4$
Bisekcioni protok = 2
Topologija prirodna za algoritme koji rade sa 1D nizovima

Dijametar – najduži među svim najkraćim putevima za svaka dva čvora

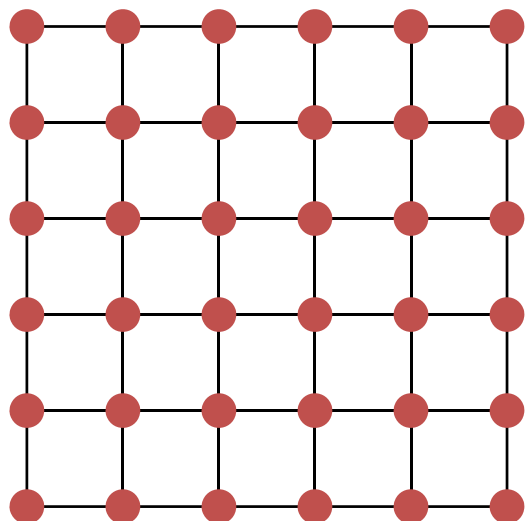
Bisekcioni protok – protok na najužem preseku neke mreže koji deli tu mrežu na dva dela



Dvodimenzioni mesh

Dijametar = $2 * (\text{sqrt}(n) - 1)$

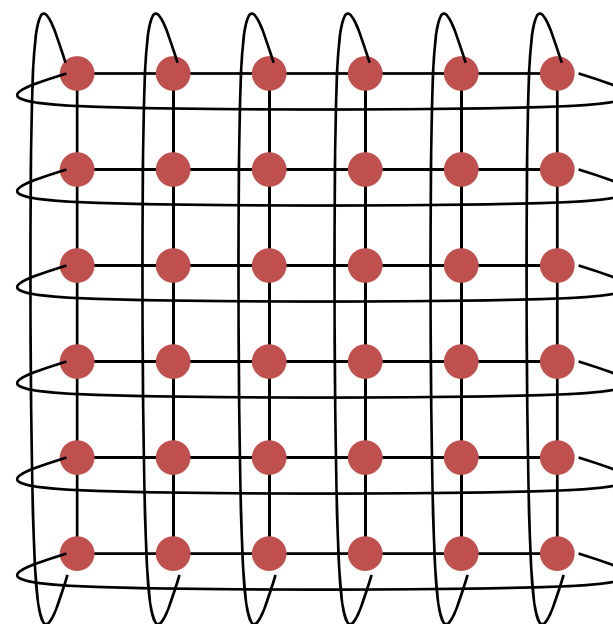
Bisekcioni protok = $\text{sqrt}(n)$



Dvodimenzioni torus

Dijametar = $\text{sqrt}(n)$

Bisekcioni protok = $2 * \text{sqrt}(n)$



Može da se generalizuje na više dimenzija (npr. Cray T3D koristi 3D Torus)
Topologija prirodna za algoritme koji rade sa 2D i/ili 3D nizovima

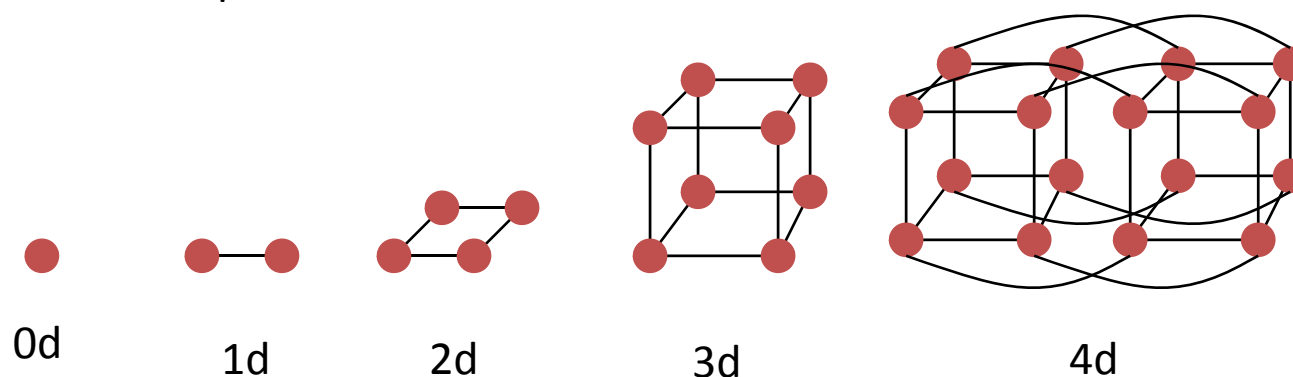


Hiperkocke

Broj čvorova $n = 2^d$

Dijametar = d

Bisekcioni protok = $n/2$

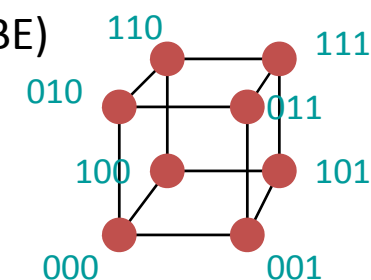


Popularni kod ranih “personalnih” superračunara (Intel iPSC, NCUBE)

Mnogo “pametnih” algoritama

Grejkod adresiranje

(svaki čvor povezan sa d drugih, a razlika samo u jednom bitu adresu)





Stabla

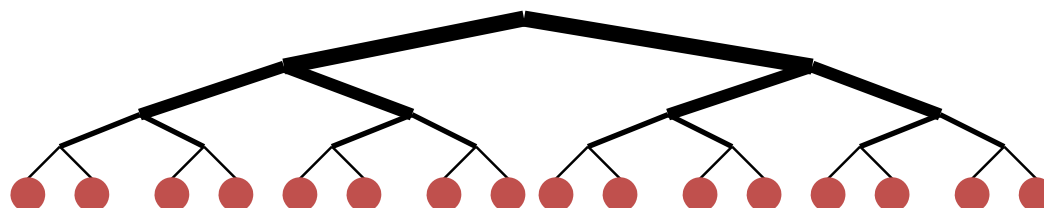
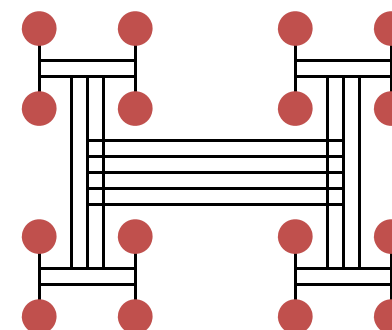
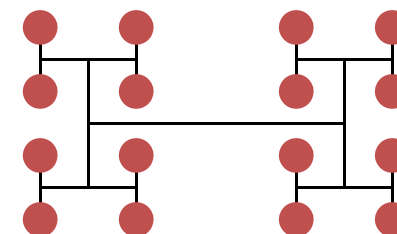
Dijametar = $\log n$

Bisekcioni protokol = 1

Lako se postavljaju kao planarni graf

Mnogi algoritmi zasnovani na stablima

Koncept "fat tree"





Butterfly sa $n = (k+1) \cdot 2^k$ čvorova

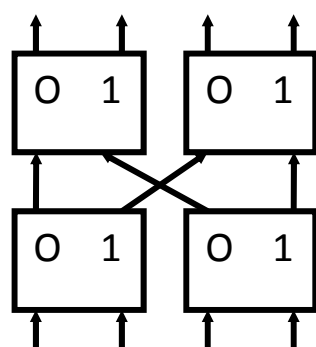
Dijametar = $2k$

Bisekcioni protok = 2^k

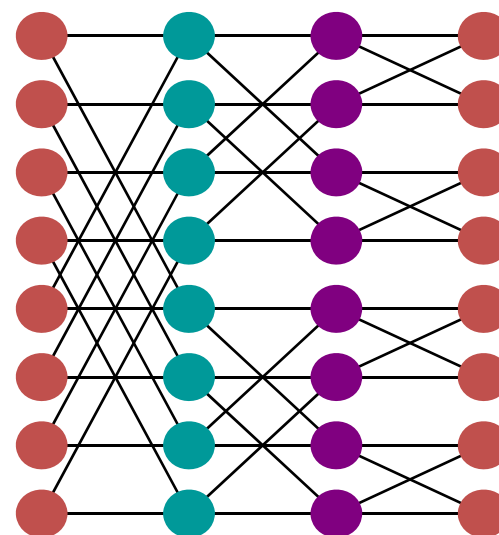
Veliki broj žica?

Topologija prirodna za FFT

Koncept višeslojne butterfly topologije



butterfly switch





Primer 9. Odrediti dostupnost servisa koji se izvršavaju na skladištu računara sa 2400 servera, ukoliko se pretpostavi da servis prestaje sa radom pri bilo kom problemu sa hardverom ili softverom. Lista mogućih problema kao i njihova učestalost pojavljivanja je data u tabeli. Pretpostaviti da popravka hardvera traje u proseku oko 1 sata.

Broj događaja godišnje	Uzrok	Posledica
2	Problem sa napajanjem	Otkaz napajanje ne obara klaster ukoliko UPS i generatori rade (generatori rade 99% vremena)
4	Obnova klastera	Planirani otkaz klastera, radi sređivanja kabliranja, obnove firmvera svičeva (switch firmware). Na svakih 9 planiranih otkaza, dolazi do jednog ne planiranog otkaza.
1000	Otkaz diska	Od 2% do 10% otkaza diskova godišnje
	Spor disk	Diskovi rade, ali 10 do 20 puta sporije
	Loša memorija	Prosečno jedan problem godišnje uzrokovan je lošim DRAM čipom
	Loše konfigurisan server	Oko 30% otkaza uzrokovano je lošom konfiguracijom servera
	Problematičan server	Oko 1% servera se rebootuje češće od jednom nedeljno
5000	Pad pojedinačnog servera	Rebutovanje servera traje oko 5 minuta

Odgovor:

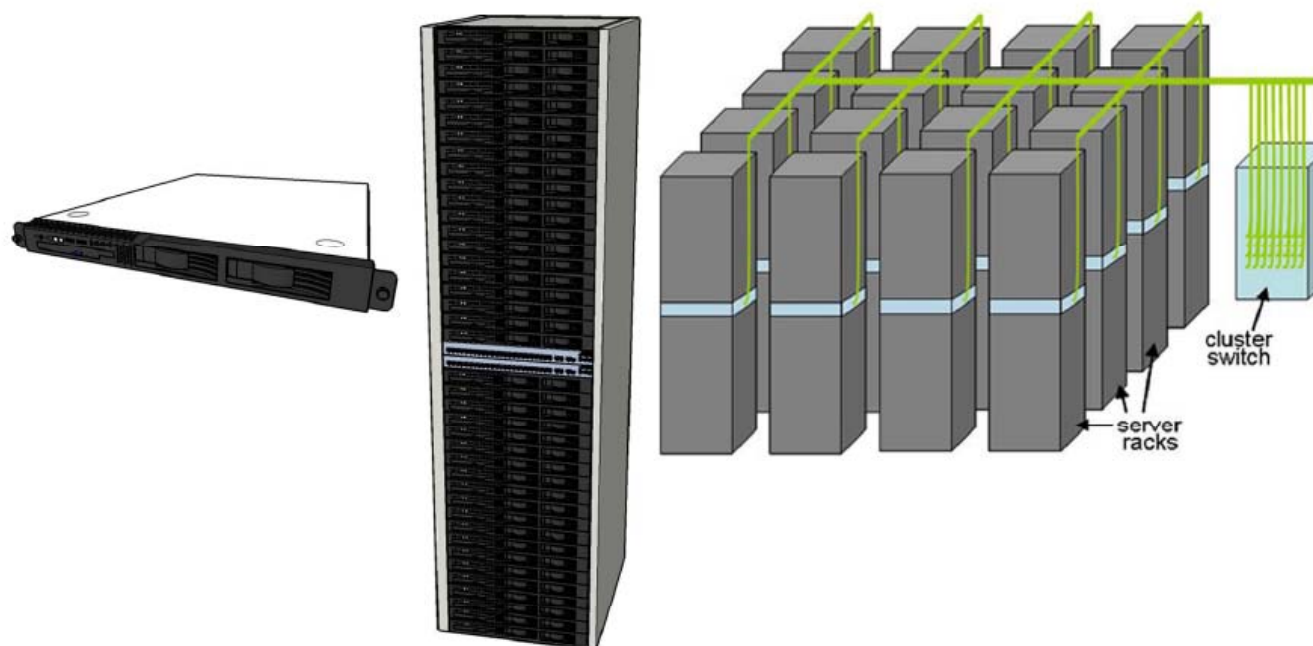
Zanemarimo: - probleme sa napajanjem, jer to isprave UPS i generatori
- probleme sa sporim diskovima jer to ne utiče na dostupnost servisa

Sati otkaza = $(4 + 250 + 250 + 250) \cdot 60 \text{ min} + (250 + 5000) \cdot 5 \text{ min} = 1191,5 \text{ sati}$

Dostupnost = $(8760 - 1191,5) / (8760) = 7568,5 / 8760 = 0,863 \rightarrow 86\%$

1191,5 sati godišnje $\rightarrow \sim 50$ dana godišnje $\rightarrow \sim$ jedan dan nedeljno

Dok za 99,9% dostupnosti (prekidi do 8,8 časova godišnje)



Server 1U – 1,75inch (4,45cm), Rack 48U – 19inch x 7foot (48,26cm x 213,36cm)

Rack switch – 48 porta, 2-8 uplink porta (oversubscription)

Cluster switch – (bisection bandwith) cena raste sa kvadratom propusne moći

	Server	Rek	Klaster
Kašnjenje DRAM (μ s)	0,1	100	300
Kašnjenje Disk (μ s)	10 000	11 000	12 000
Propusna moć DRAM (MB/s)	20 000	100	10
Propusna moć Disk (MB/s)	200	100	10
Kapacitet DRAM (GB)	16	1040	31 200
Kapacitet Disk (GB)	2000	160 000	4 800 000



Primer 10. Karakteristike hijerarhije RAM memorije u skladištu računara su date u tabeli.

	Server	Rek	Klaster
Kašnjenje DRAM (μ s)	0,1	100	300
Kašnjenje Disk (μ s)	10 000	11 000	12 000
Propusna moć DRAM (MB/s)	20 000	100	10
Propusna moć Disk (MB/s)	200	100	10
Kapacitet DRAM (GB)	16	1040	31 200
Kapacitet Disk (GB)	2000	160 000	4 800 000

a) Koje je prosečno kašnjenje memorije ukoliko je poznato da je 90% pristupa unutar servera, 9% van servera, ali unutar rek ormana, dok je preostalih 1% van rek ormana, ali unutar klastera?

Odgovor: $(90\% \cdot 0,1) + (9\% \cdot 100) + (1\% \cdot 300) = 12,09 \mu\text{s}$

→ 120 puta sporije nego da je svih 100% pristupa bilo lokalno unutar servera

b) Koliko traje transfer 1000MB podataka između diskova na istom serveru, između diskova na različitim serverima unutar rek ormana, između diskova na različitim serverima koji se nalaze u različitim rek ormanima istog klastera? Koliko traje prenos iste količine podataka u memoriji u sva tri navedene slučajeve?

Odgovor: Unutar servera: → disk ($1000/200 = 5\text{s}$)

memorija ($1000/20000 = 0,05\text{s}$)

Unutar reka: → disk ($1000/100 = 10\text{s}$)

memorija ($1000/100 = 10\text{s}$)

Unutar klastera: → disk ($1000/10 = 100\text{s}$)

memorija ($1000/10 = 100\text{s}$)



Jednostavan model vremena izvršavanja posla u klasteru računara

Broj jezgara u jednom čvoru klastera – m

Vreme pristupa deljenoj DRAM memoriji $\sim 100\text{ns}$

Broj čvorova u klasteru – n

Vreme pristupa globalnim pristupima između čvorova $\sim 100\ \mu\text{s}$ (pristup preko LAN)

Fiksno vreme izračunavanja na jednom čvoru – 1ms

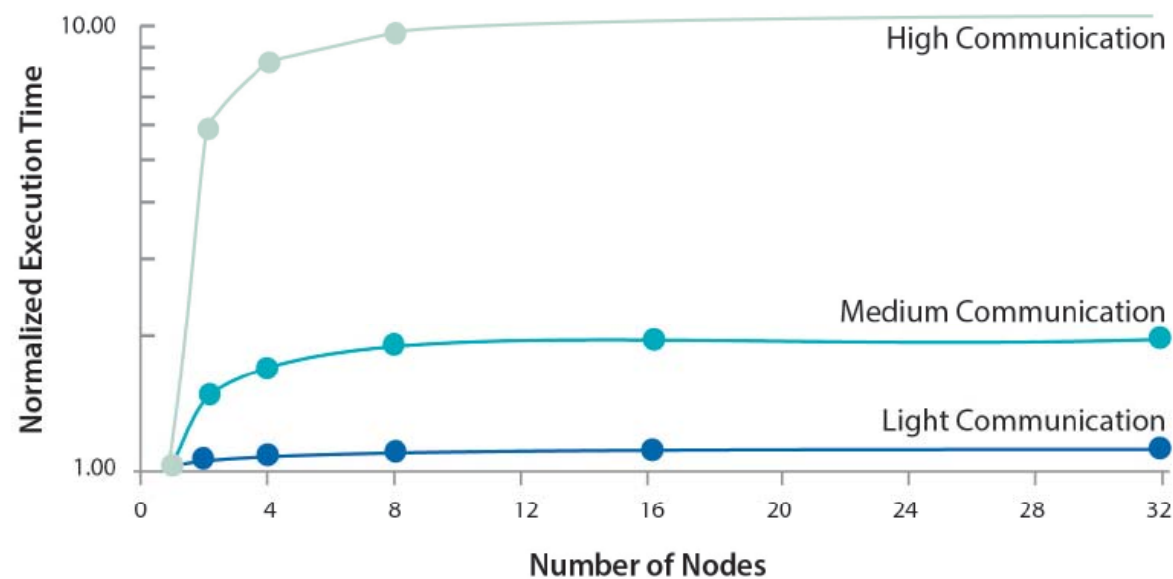
Globalni pristupi uniformno raspoređeni

Količina globalnih pristupa koja se mapira na lokalne pristupe
je obrnuto proporcionalna broju čvorova

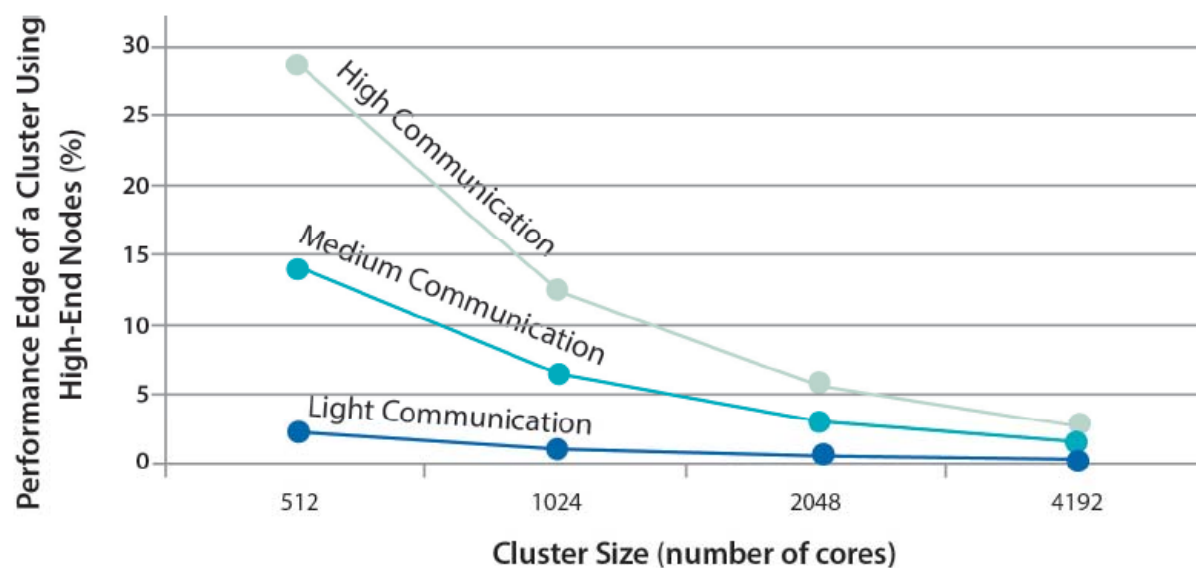
Vreme izvršavanja = $1\text{ms} + f \cdot (100\text{ns}/n + 100\mu\text{s} \cdot (1-1/n))$

f – broj globalnih pristupa po jedinici posla (na svaku 1ms posla)

($f=1$ – slaba komunikacija; $f=10$ – umerena komunikacija; $f=100$ – jaka komunikacija)



Da li koristiti jedan server sa 128 jezgara ili trideset dva servera sa po 4 jezgara?





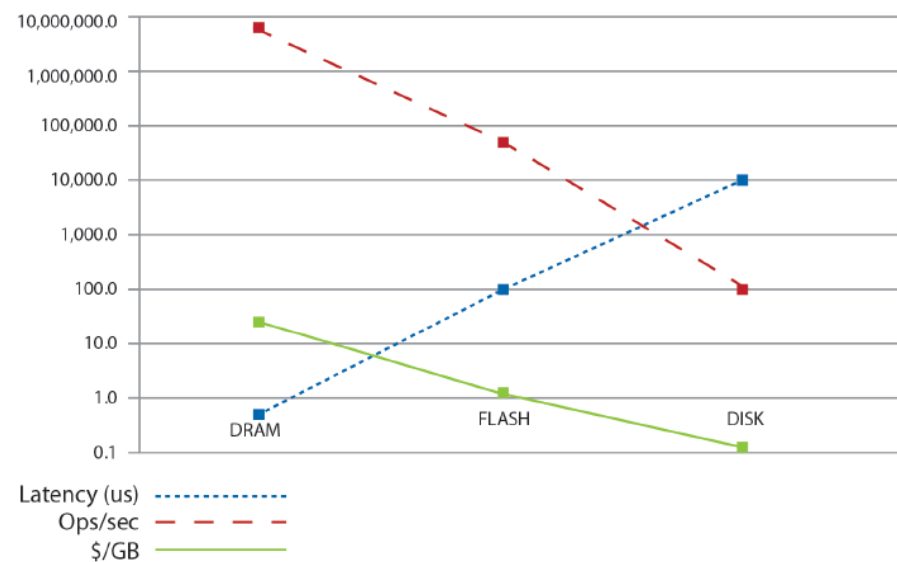
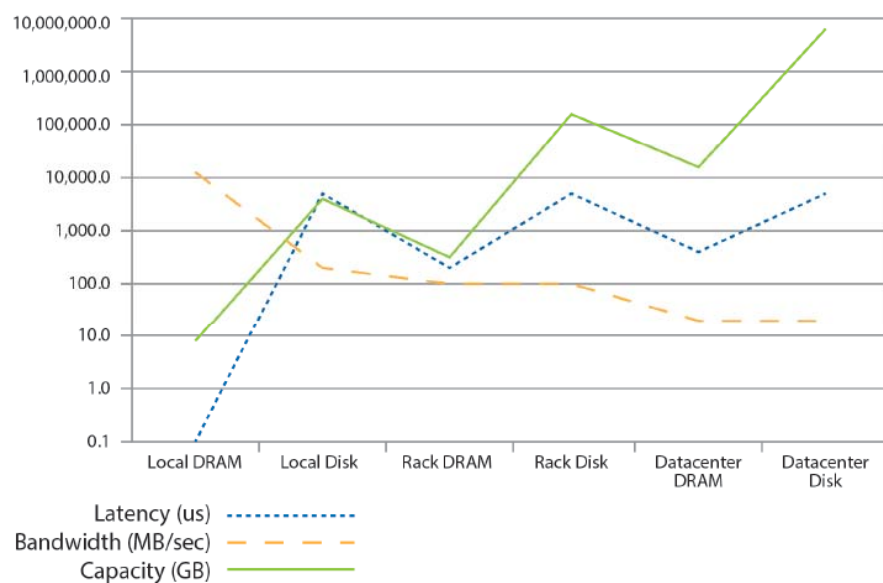
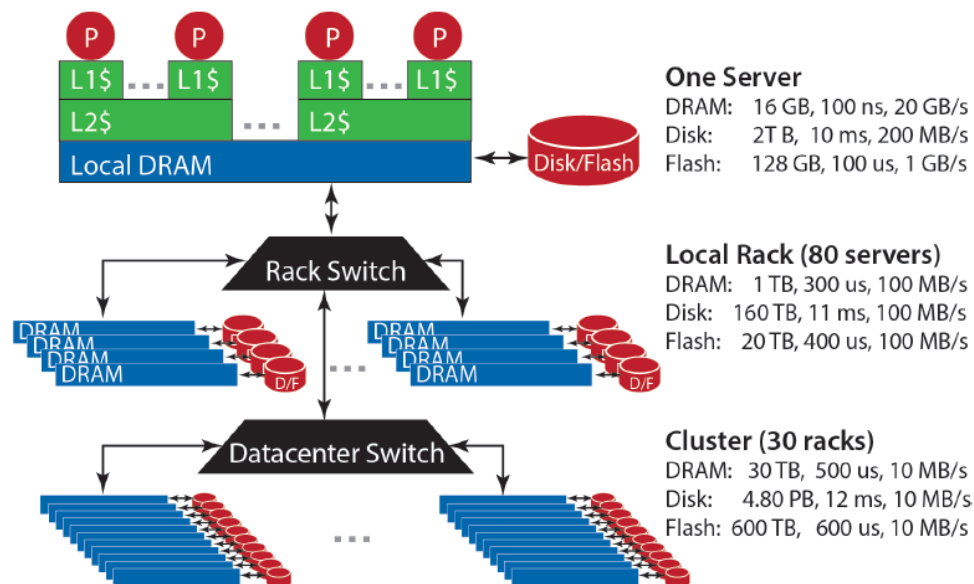
Da li koristiti “lošije” procesore za izgradnju klastera?

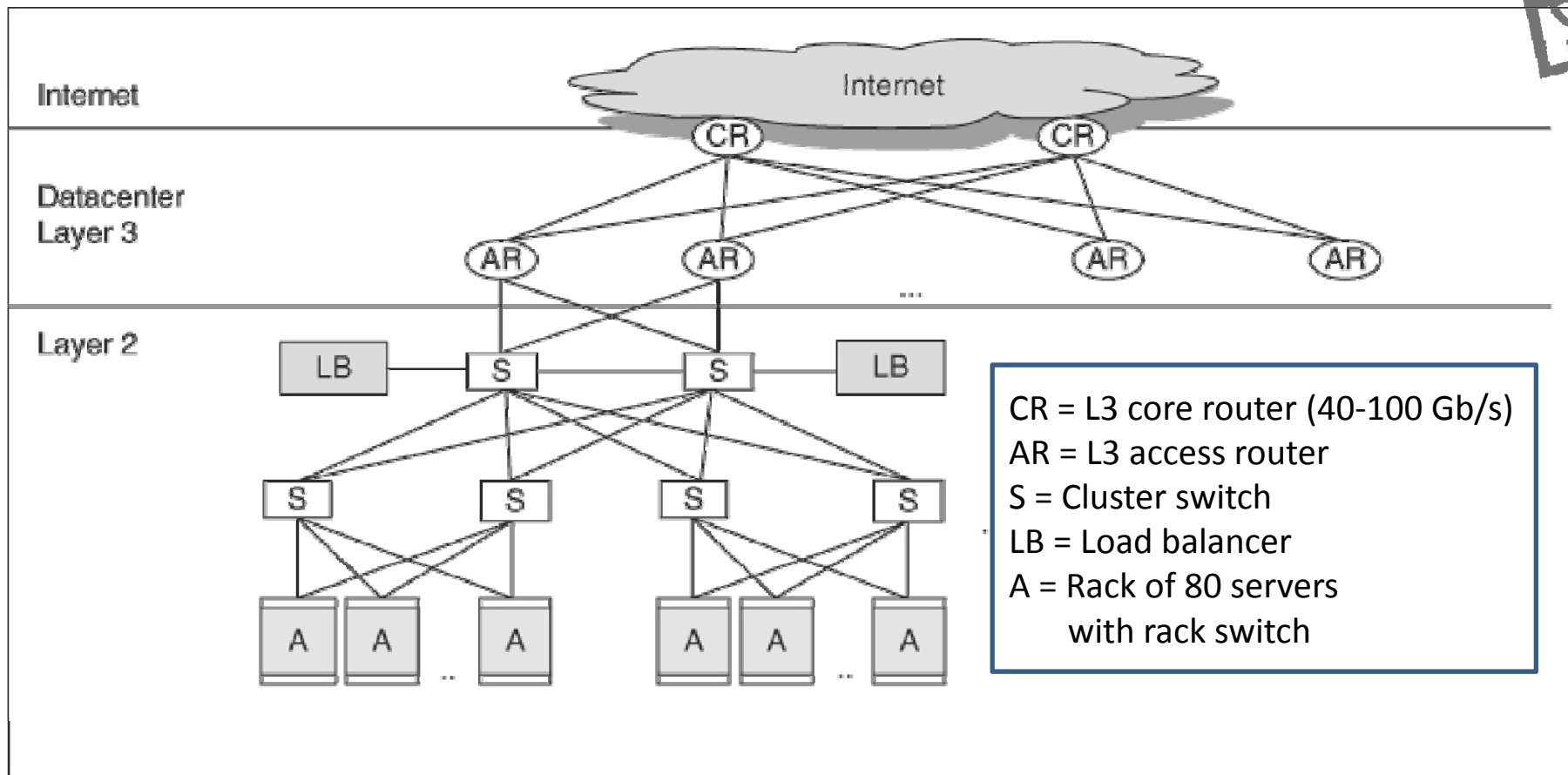
Prednosti:

- “lošiji” procesori imaju bolji odnos cena/performanse
- većina posla ima mnogo I/O pristupe, tako da se “bolji” procesori loše skaliraju
- “lošiji” procesori su efikasniji u pogledu potrošnje (manja učestanost takta i niži napon)

Mene:

- potrebna složenija mrežna infrastruktura
- veća paralelizacija donosi:
 - kompleksnije algoritme
 - nepredvidivost u performansama
 - veća globalna komunikacija (veća potreba za sinhronizacijom manjih poslova)
- složeniji razvoj softvera
- niži stepen iskorišćenosti (problem konsolidacije - ne može više VM na jednom čvoru)





Server: faktor 2U, dva procesora, 1Gb/s LAN

16GB RAM sa 100ns pristupom i transferom 20GB/s

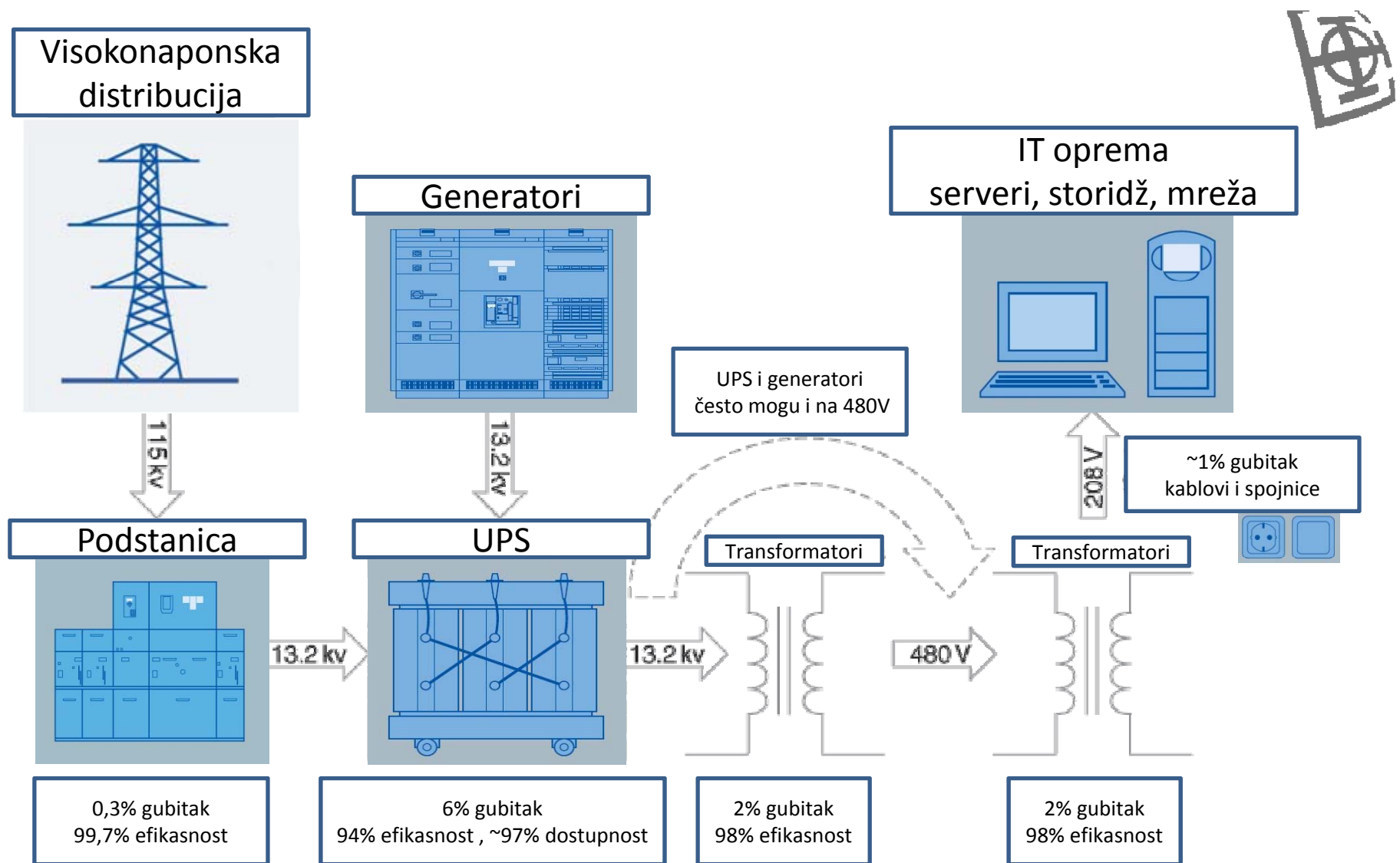
2TB HD sa transferom od 200MB/s

Rek : 80 servera, 1TB RAM i transferom 100MB/s, 160TB HD sa transferom 100MB/s

Klaster: 30 rek ormana, 30TB RAM i transferom 10MB/s, 4,8PB HD sa transferom 10MB/s

Google - <http://www.youtube.com/watch?v=zRwPSFpLX8I>

Microsoft - <http://www.youtube.com/watch?v=hOxA1l1pQlw>

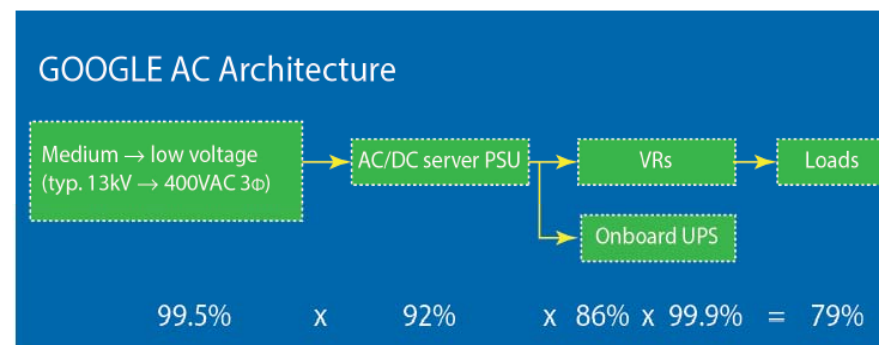
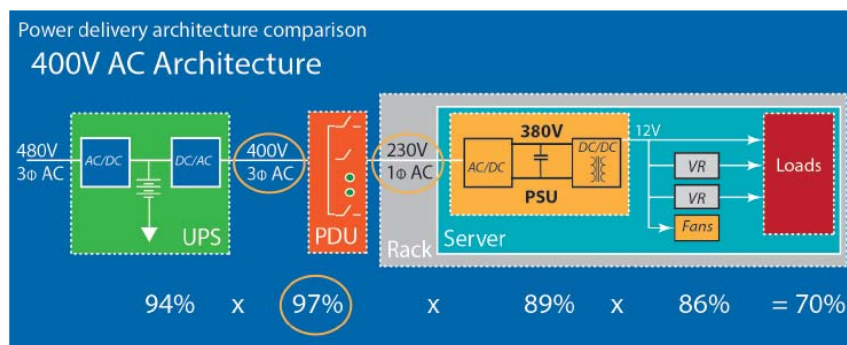
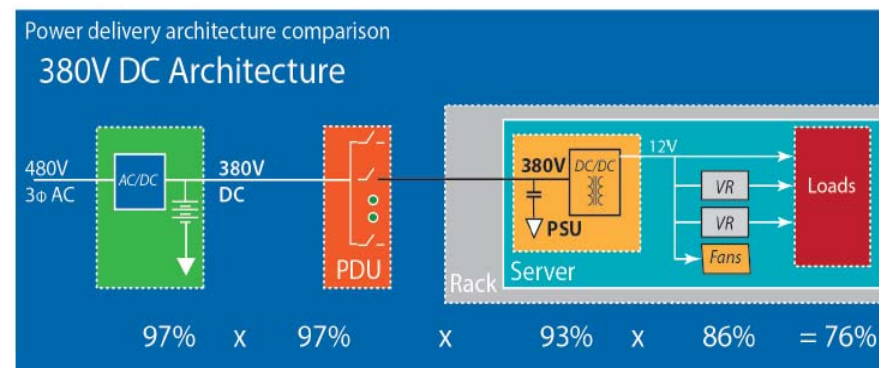
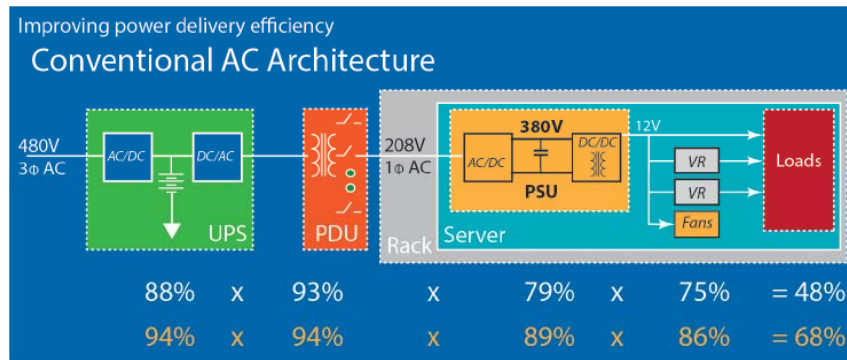


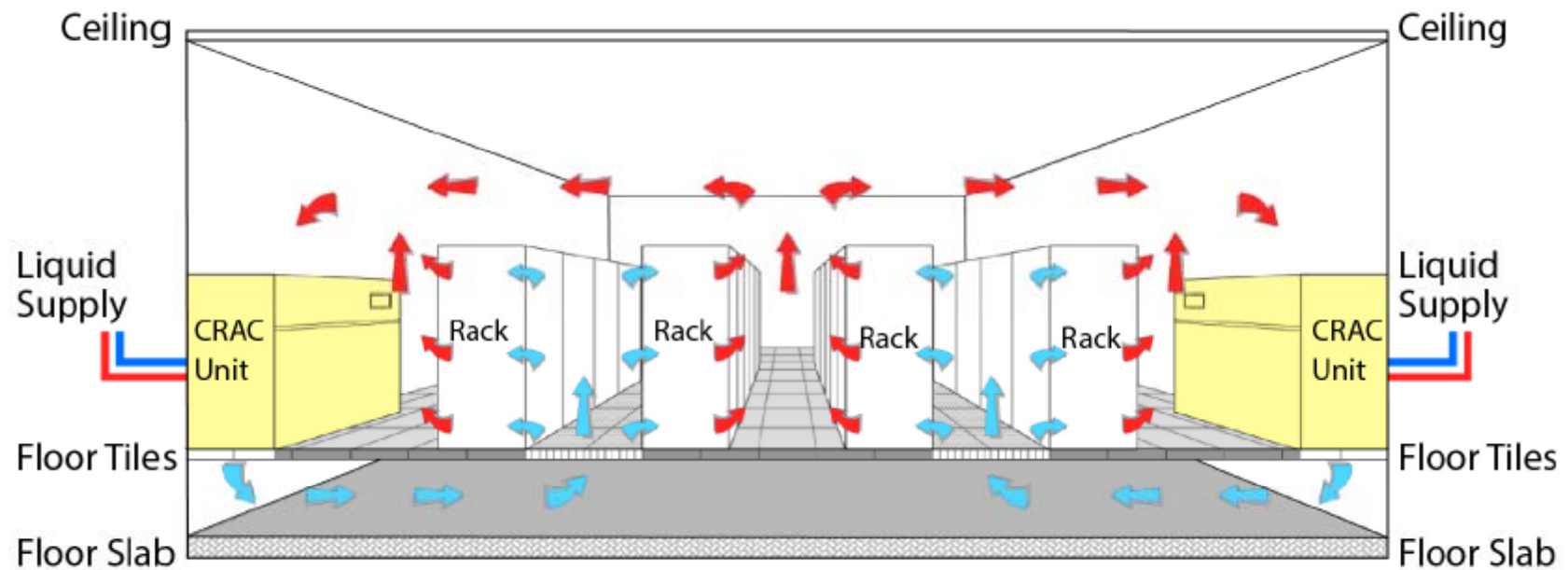
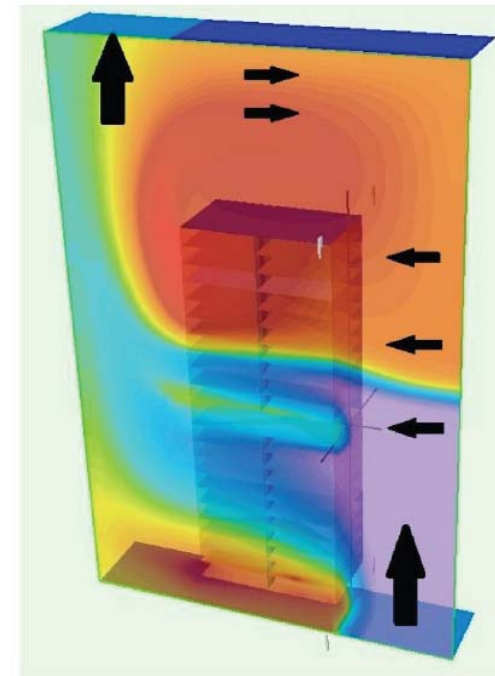
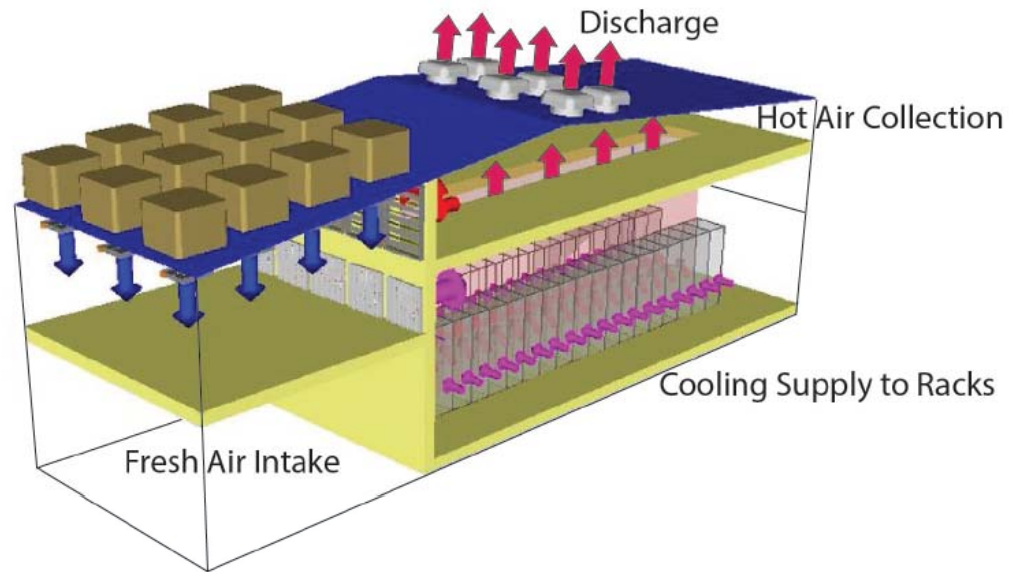
Efikasnost (u "idealnom" slučaju): $0,997 \cdot 0,94 \cdot 0,98 \cdot 0,98 \cdot 0,99 = 0.891 \sim 89\%$

UPS ~ 7%-12% cene celokupne IT opreme

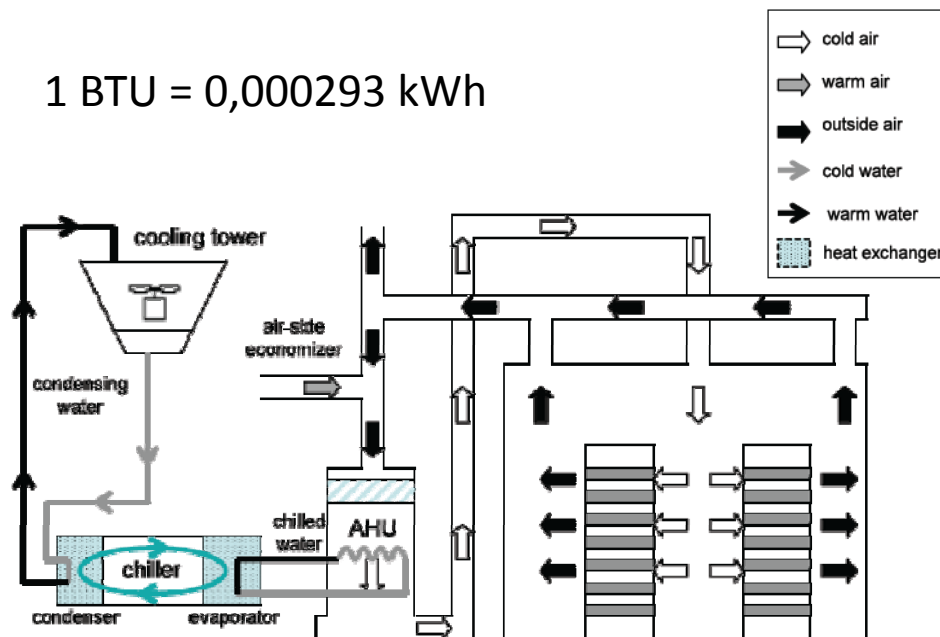
PDU ~ 75-225KW → ~ 10 rek ormana

Efikasnost napajanja servera, tj. napajanja sa više nivoa izlaznog napona (max. 87%-90%)

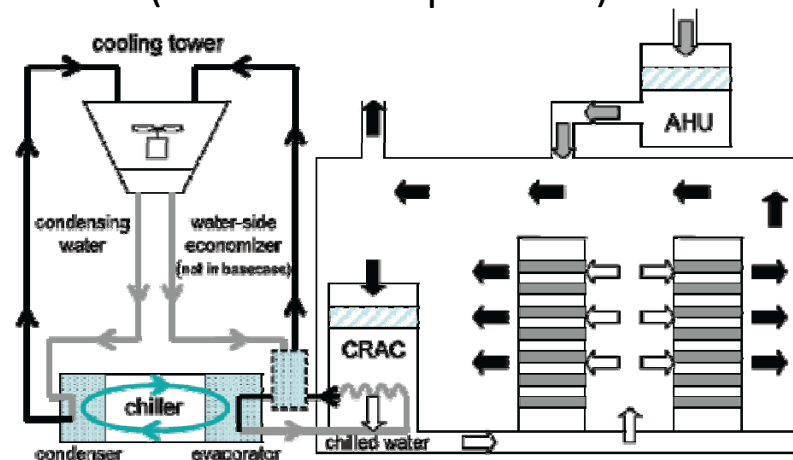




1 BTU = 0,000293 kWh



Temperatura
vlažnog termometra
(wet-bulb temperature)



CRAC (Computer Room Air-Conditioning) – temperatura vazduha kod servera je 18-22°C

CRAC: 10-20% , Chillers: 30-50% snage celokupne IT opreme

Unutar IT opreme: 33% procesori, 30% memorija, 10% diskovi, 5% mreža, 22% ostalo

Efikasnost skladišta računara:

PUE (Power Utilization Effectiveness) = Ukupna snaga skladišta / Snaga IT opreme

PUE > 1 (npr. za PUE od 1,69 čak 0,55 otpada na hlađenje)

(veoma neefikasno-3; neefikasno-2,5; prosečno-2; efikasno-1,5; veoma efikasno-1,2)

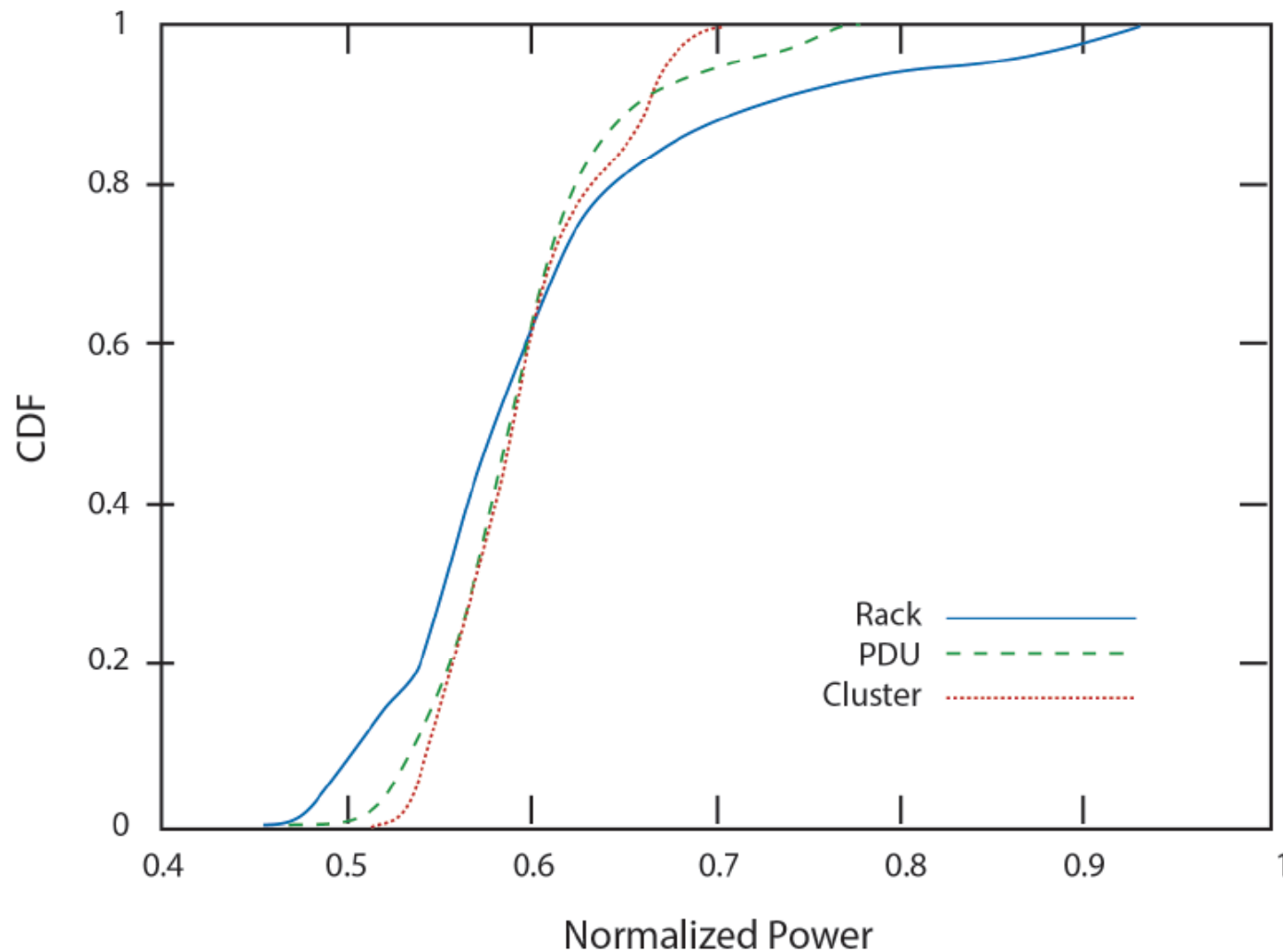
Klasifikacija :

Nivo 1: - po jedna putanja za distribuciju napajanja i hlađenja

Nivo 2: - uvođenje redundantnih komponenti (tzv. N+1 dizajn)

Nivo 3: - više putanja, jedna aktivna, mogućnost paralelnog održavanja (tzv. N+2 dizajn)

Nivo 4: - nema jedinstvenu tačku otkaza (npr. dupla aktivne putanje)



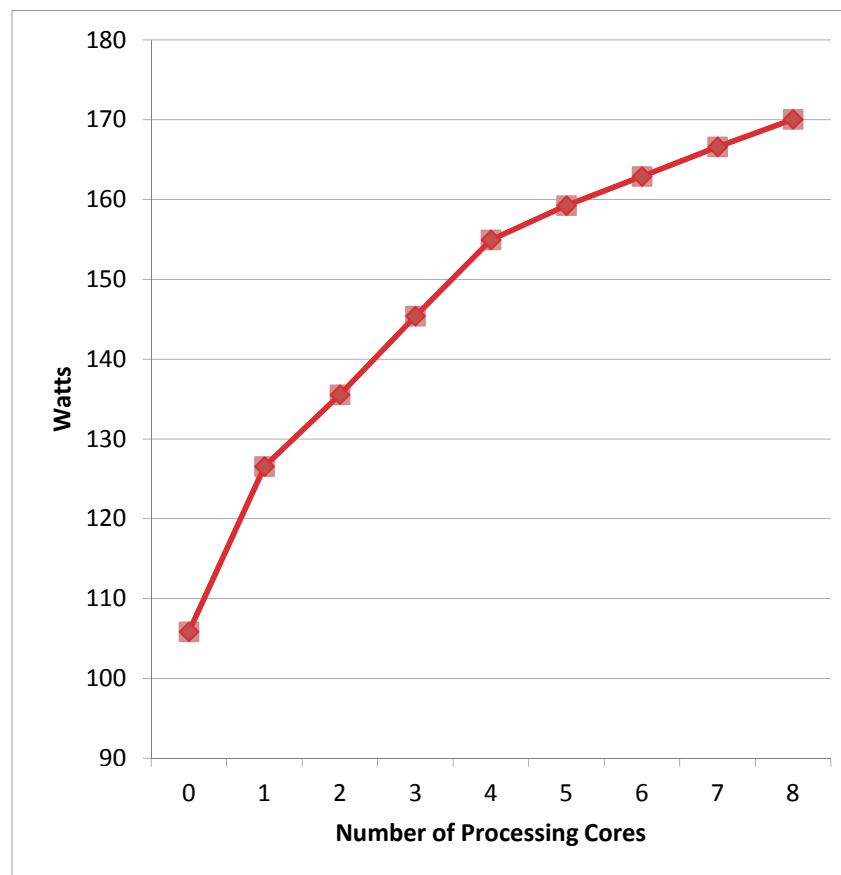
CDF – kumulativna distribucija upotrebe snage

Rack (80 servera), PDU (800 servera), Cluster (5000 servera)

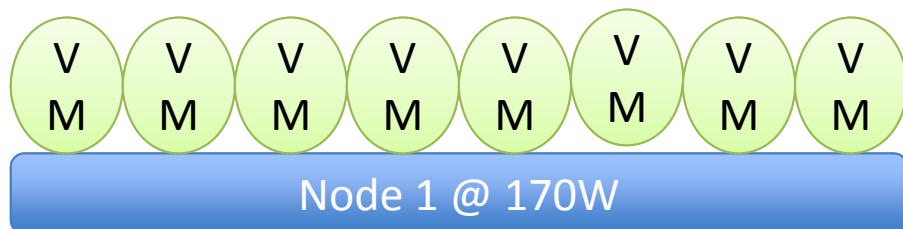
Rack je 80% vremena ispod 65% upotrebe snage

Cluster nikada nije iskoristio više od 72% od svoje ukupne agregirane snage

→ iskoristiti za oversubscription snage



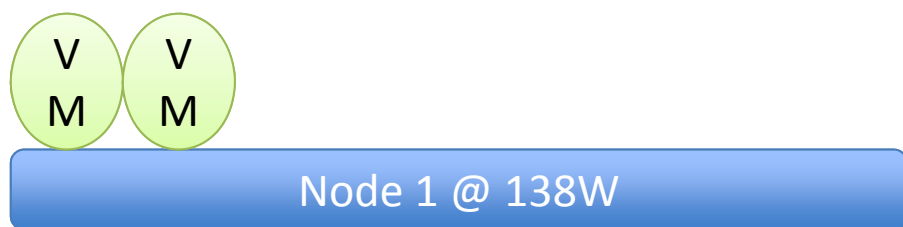
Potrošnja električne energije za
Intel Core i7 920 Server
(4 jezgra, 8 virtuelnih jezgra)



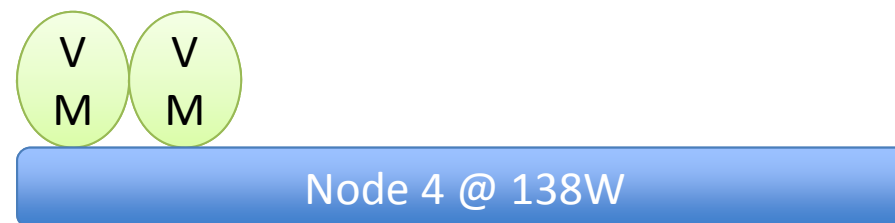
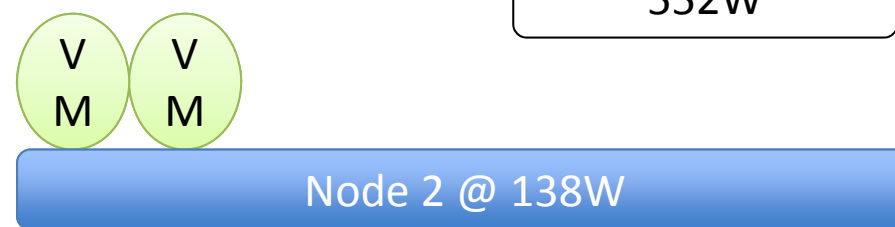
485W

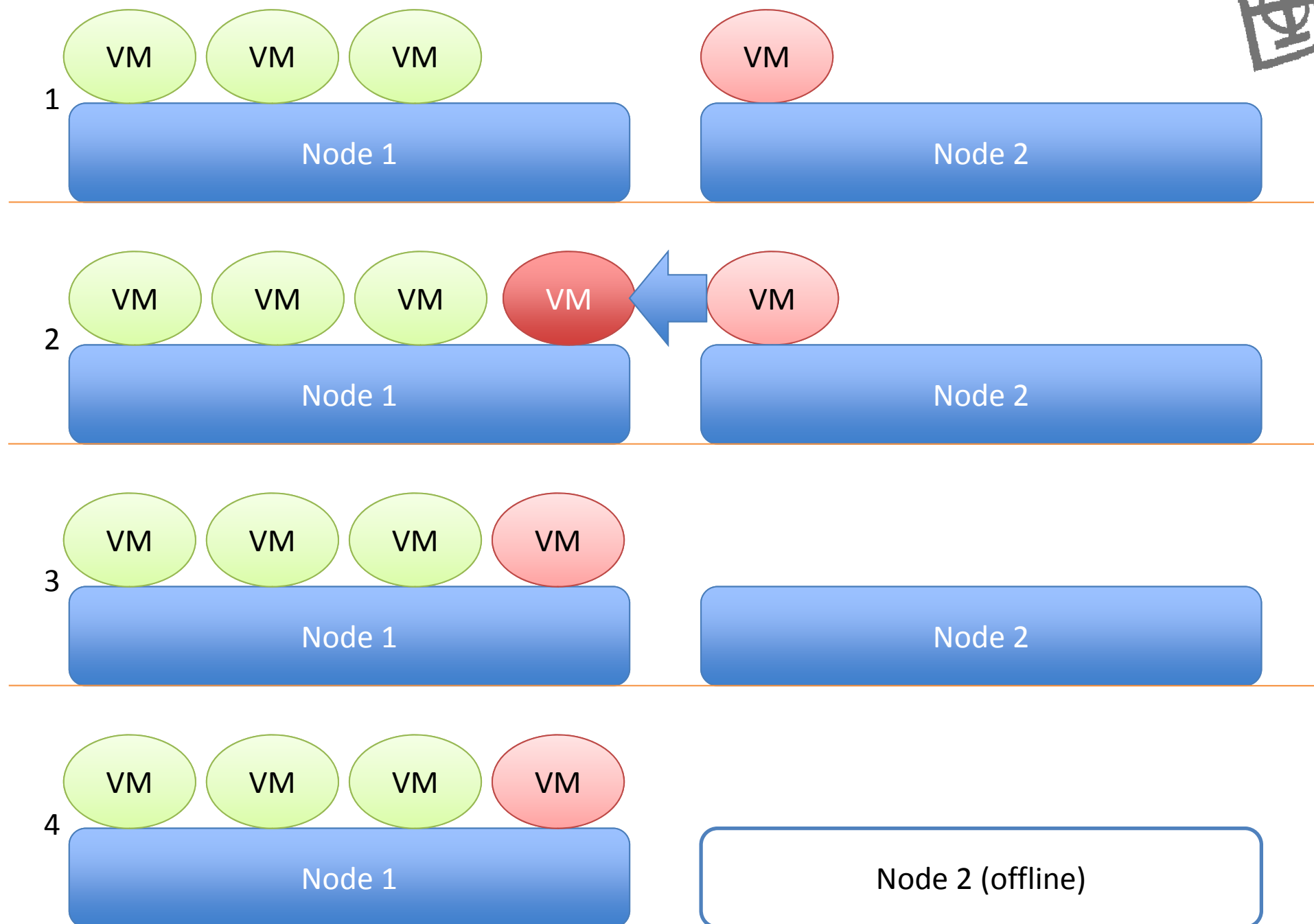


VS.



552W







→ Centri srednje veličine (1000 servera) vs. velikih centara (50.000 servera)

→ Tehnologija	srednji centri	veliki centri	odnos
Mreža	\$95 per Mbit/sec/month	\$13 per Mbit/sec/month	7,3
Storidž	\$2.20 per GByte / month	\$0.40 per GByte / month	5,5
Administracija	140 Servers / Administrator	>1000 Servers / Administrator	7,1

→ Cena električne energije po regijama (SAD)

→ Cena KWH	Gde	Zašto
3.6¢	Ajdaho	Hidroenergija;
10.0¢	Kalifornija	ne šalje se na veliku udaljenost;
		šalje se na veliku udaljenost;
		kapacitet mreže u regionu;
		termoelektrane nisu dozvoljene.
18.0¢	Havaji	gorivo se mora transportovati.



Primer 11. Potrebno je odrediti kapitalne troškove (CAPEX – Capital Expenditure) izgradnje skladišta računara sa oko 50 000 servera na lokaciji koja može da proizvede 11,6MW snage. Po ceni od 9\$ po kritičnom W snage skladišta, definiše se 81,8% cene izgradnje skladišta, dok je vrednost za PUE jednaka 1,45. Popust na aktivnu mrežnu opremu je 60%.

Date su sledeće pretpostavke po pitanju IT opreme:

Server → 1U, potrošnja 165W, cena 1450\$

Rack → 40 servera

Core router → potrošnja 5KW, cena 500 000\$

Border router → potrošnja 5KW, cena 362 000\$

Aggregation switch → potrošnja 10KW, 128 port, cena 750 000\$, dovoljan za 54 reka

Access switch → potrošnja 151W, 48 port, cena 11,995\$

Odgovor:

Kritična snaga = Ukupna snaga / PUE = 11,6MW/1,45 = 8MW

Pretpostavka je da mrežna oprema ~5% snage → za servere ~ 7,6MW

Procenjen broj servera: 7,6MW / 165W → ~ 46060 servera → ~1152 rek ormana

→ ~1152 rek sviča → ~22 klaster sviča

Procena potrošnje mrežne opreme: $1152 \cdot 151W + 22 \cdot 10000 + 2 \cdot 5000 + 2 \cdot 5000 = 413952W$

$7586048W / 165W \rightarrow 45976$ servera → 1150 rek ormana → 413650W → 45978 servera

CAPEX = $45978 \cdot 1450 + (1150 \cdot 11995 + 22 \cdot 750K + 2 \cdot 362K + 2 \cdot 500K) \cdot 0,4 + 8M \cdot 9\$ / 0,818$

CAPEX = serveri + mreža+ skladište = 66,70M + 12,81M + 88M = 167,51M US\$



Primer 12. Potrebno je odrediti operativne troškove (OPEX – Operational Expenditure) skladišta računara iz primera 11. Cena struje je 0,07\$ po KWh. Cena kredita sa fiksnom kamatnom stopom i ratom je 5%. Period amortizacije skladišta je 10 godina, mrežne opreme 4 godine, a IT opreme 3 godine. Prosečno iskorišćenje IT opreme je 80%. Plate zaposlenih su 85000US\$ mesečno.

Odgovor:

CAPEX = serveri + mreža+ skladište = 66,70M + 12,81M + 88M = 167,51M US\$

Amortizacija

Mesečna otplata: $C = K \cdot P_0 / (1 - (1 + K)^{-N})$

K – mesečna kamatna stopa, P_0 – iznos pozajmice, N – broj rata

Mesečna amortizacija servera = $((5/100/12) / (1 - (1 + 5/100/12)^{-3 \cdot 12})) \cdot 66\,700\,000$

→

Mesečna amortizacija servera = 1 998 097

Mesečna amortizacija mreže= 294 943

→

Cena opreme za infrastrukturu = 88 000 000 · 0,818= 72M

Mesečna amortizacija opreme za infrastrukturu= 763 672

Mesečni troškovi struje= $((11,6\text{MW} \cdot 0,8) \cdot 1000) \cdot 0,07 \cdot (24 \cdot 365/12) = 474\,208$

OPEX= 1 998 097 + 294 943 + 763 672 + 474 208 + 85 000 = 3 615 920 ~3,6M US\$



Mesečna otplata: $C = K \cdot P_0 / (1 - (1+K)^{-N})$

K – mesečna kamatna stopa, P_0 – iznos pozajmice, N – broj rata

P_0 – iznos pozajmice

$P_1 = P_0 + P_0 \cdot K - C = P_0(1+K) - C$ – iznos pozajmice nakon otplate prve rate

$P_2 = P_1(1+K) - C = P_0(1+K)^2 - C(1+K) - C$

...

$P_N = P_0(1+K)^N - C((1+K)^{N-1} + (1+K)^{N-2} \dots + 1) = P_0(1+K)^N - C \cdot (S)$

$S = (1+K)^{N-1} + (1+K)^{N-2} \dots + 1 \rightarrow S \cdot (1+K) = (1+K)^N + (1+K)^{N-1} \dots + (1+K)$

$\rightarrow S \cdot (1+K) - S = (1+K)^N - 1 \rightarrow S = ((1+K)^N - 1) / K$

$\rightarrow P_N = P_0(1+K)^N - C \cdot (S) = P_0(1+K)^N - C \cdot (((1+K)^N - 1) / K)$

Na kraju otplate: $P_N = 0 \rightarrow C = K \cdot P_0 / (1 - (1+K)^{-N})$



CAPEX = 167,51M

OPEX = 3,6M

→ OPEX = 2,15% CAPEX

Plate zaposlenih = 2,36% OPEX

Serveri →	~40% OPEX	~55% CAPEX
Mrežna infrastruktura →	~8% OPEX	~7,5% CAPEX
Infrastruktura skladišta →	~43% OPEX	~21% CAPEX

Cena po vatu snage godišnje = $12 \cdot (763\,672 + 474\,208) / 8\,000\,000 = 1.86$ US \$
(Mesečna amortizacija opreme za infrastrukturu + Mesečni troškovi struje) / (Kritična snaga)

Cena po server satu = $\text{OPEX} / \text{Broj_servera} / \text{Broj_sati_mesečno}$
 $= 3,6\text{M US\$} / 45978 / 730 = 0,107$ US \$/h



Primer 13. Posmatra se skladišta računara iz primera 11. Kako se menja CAPEX ukoliko se predviđeni serveri zamene novim koji su 10% bržim pri istom iskorišćenju? Cena bržih servera je za 20% veća. Kako se menja OPEX ukoliko brži serveri imaju 15% veću potrošnju?

Odgovor:

Novi broj servera: $N_{\text{novo}} \cdot 1,1 = N_{\text{staro}} \rightarrow N_{\text{novo}} = 45978 / 1,1 = 41799$ servera

Cena novih servera: $1450\$ \cdot 1,2 = 1740\$ \rightarrow$ ukupno: $41799 \cdot 1740\$ = 72730260\$$

Potrošnja: $165W \cdot 1,15 = 189.75 \sim 190W \rightarrow 41799 \cdot 190W = 7941810W$

41799 servera $\rightarrow \sim 1045$ rek ormana

Procena potrošnje mrežne opreme: $1045 \cdot 151W + 20 \cdot 10000 + 2 \cdot 5000 + 2 \cdot 5000 = 377795W$

\rightarrow Kritična snaga: $7941810W + 377795W = 8319605W$

PUE = 1,45 \rightarrow Potrebna snaga: $8319605W \cdot 1,45 = 12063427.25W$

\rightarrow Mesečni troškovi struje = $(12063427.25 \cdot 0,8 / 1000) \cdot 0,07 \cdot (24 \cdot 365 / 12) = 493\ 153$

Mesečni troškovi struje su uvećani za $493\ 153 - 474\ 208 = 18945\$$



Softverska infrastruktura

Softver na nivou aplikacije

Paketna obrada, Interaktivni servisi

Softver na nivou klastera

Upravljanje resursima
Apstrakcija hardvera
(GFS, Hadoop, Dynamo, Chubby)
Postavljanje i održavanje
(Autopilot, X-Trace, System Health)
Programerska okruženja
(Map Reduce, BigTable , Dynamo)

Softver na nivou platforme

OS, Standardne biblioteke

	Performanse	Dostupnost
Replikacija	Da	Da
Particionisanje	Da	Da
Balansiranje opterećenja	Da	
Nadgledanje sistema		Da
Provera integriteta		Da
Kompresija na nivou aplikacije	Da	
Eventualna konzistentnost	Da	Da



Centralizovano nadgledanje?

→ olakšava rad u distribuiranom sistemu

→ kako rešiti problem jedne tačke otkaza i uskog grla po pitanju performansi

Reed-Slomon kodovi?

→ povećava dostupnost → manji utrošak prostora nego pri replikaciji

Canaries?

→ povećava dostupnost → pokretanje distribuirane obrade, prvo na manjoj grupi servera

Tail-tolerance?

→ povećava performanse

→ redundantno izvršavanje poslova (subtasks)

