Optimalno 11g RAC skaliranje na HP blade poslužiteljima

Draško Tomić, Tomislav Lukačević
Hewlett-Packard

Agenda

Oracle RAC skaliranje

Oracle 11g RAC Grid svojstva

Pregled HP blade sustava

Optimalna implementacija Oracle 11g RAC na HP blade poslužiteljima

Demo

Oracle RAC skaliranje

Statički scale-up/scale-out

- Statički scale-up
  - Označava povećanje resursa nekog poslužitelja (procesori, memorija, I/O)
  - Uobičajeno scale-up rješenje je UNIX server sa 2 – 128 procesora
  - Ovakvi sustavi donose smanjenu latenciju memorije i šire područje adresiranja za bolje performanse baza podataka

- Statički scale-out
  - Odnosi se na dodavanje novih poslužitelja u postojeću infrastrukturu, kako bi se opterećenje proširilo na više sustava
  - Uobičajeno scale-out rješenje sadrži dva ili više Linux poslužitelja
  - Rješenje podržava modularni porast performansi dodavanjem novih poslužitelja

Dinamički scale-up/scale-out

- Dinamički scale-up
  - Proširenje resursa (procesora, memorije, I/O modula) neke particije bez prekida u radu iste
  - Uobičajene tehnologije su OL* ili vPar

- Dinamički scale-out
  - Aktivacija neaktivnih poslužitelja u nekom RAC klasteru, kako bi se opterećenje raspodijelilo na više poslužitelja
  - Već aktivni poslužitelji mogu odrađivati neki drugi posao prije ili za vrijeme dinamičkog širenja
  - Učinkovit način pridjeljivanja računarskih resursa aplikacijama sa relativno velikim fluktuacijama u opterećenju

Važnije scale-up značajke

- Jednostavno upravljanje (manje poslužitelja)
- Bolje performanse (backplane brži od prospojne mreže)
- Finija granularnost (iCAP i TiCAP na nivou komponenti pojedinog računala)

Važnije scale-out značajke

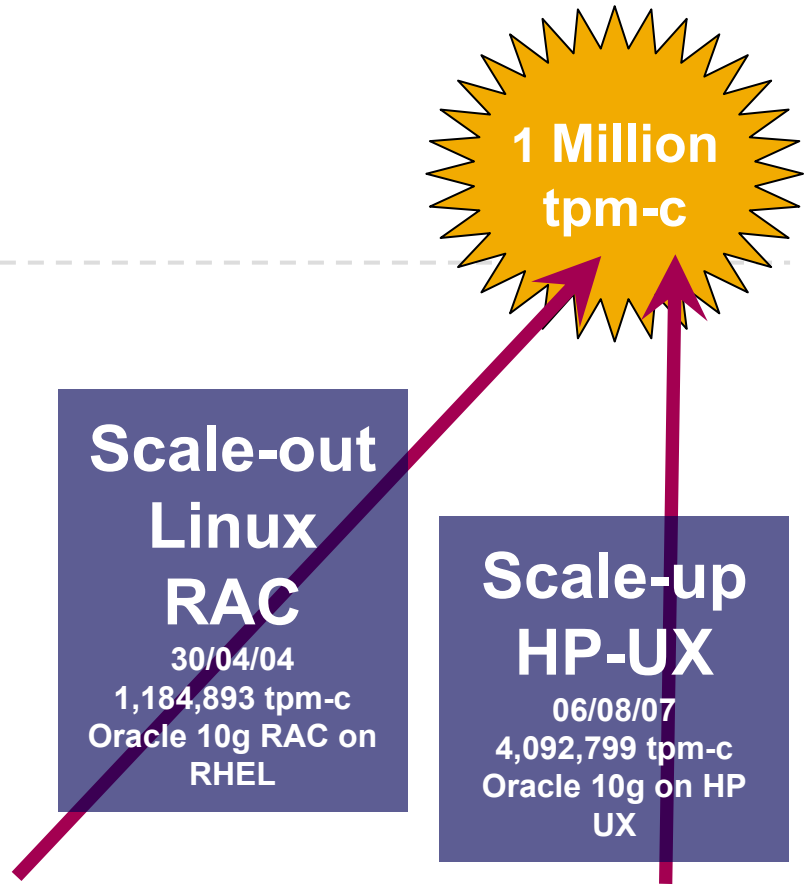
- Niža cijena hardvera(jedno računalo sa N*M procesora je skuplje od N računala sa po M procesora svaki)
- Proširivost (teoretski neograničena)
- Veća raspoloživost i pouzdanost

Preuzimanje: modeli i intervali

- Single-instance Oracle
  - active/standby model
  - intervali preuzimanja od 5 do 20 minuta, ovise o:
    - Veličini baze podataka, # LUNs/VGs/DGs
    - Checkpoint intervalu
    - Heartbeat intervalu i Node Timeout vrijednostima
- Oracle RAC
  - dijeljeni (active/active) model
  - intervali preuzimanja od 20 – 60 sec., ovise o:
    - Checkpoint intervalu
    - Upotrebi SGeff (Serviceguard extension for faster failover)

www.tpc.org

- RHEL/RAC 10g: 1,184,893 TPM
  - 16 x 4-way poslužitelji
  - 64 x 1.5Ghz Itanium single-core
  - 40 x 1 Gb Eth. ports
- HP-UX/10g: 1,008,144 TPM
  - Superdome 64-way
  - 64 x 1.5Ghz Itanium single-core
- HP-UX/10g: 4,092,799 TPM
  - Superdome 128-way
  - 64 x 1.6Ghz Itanium dual-core

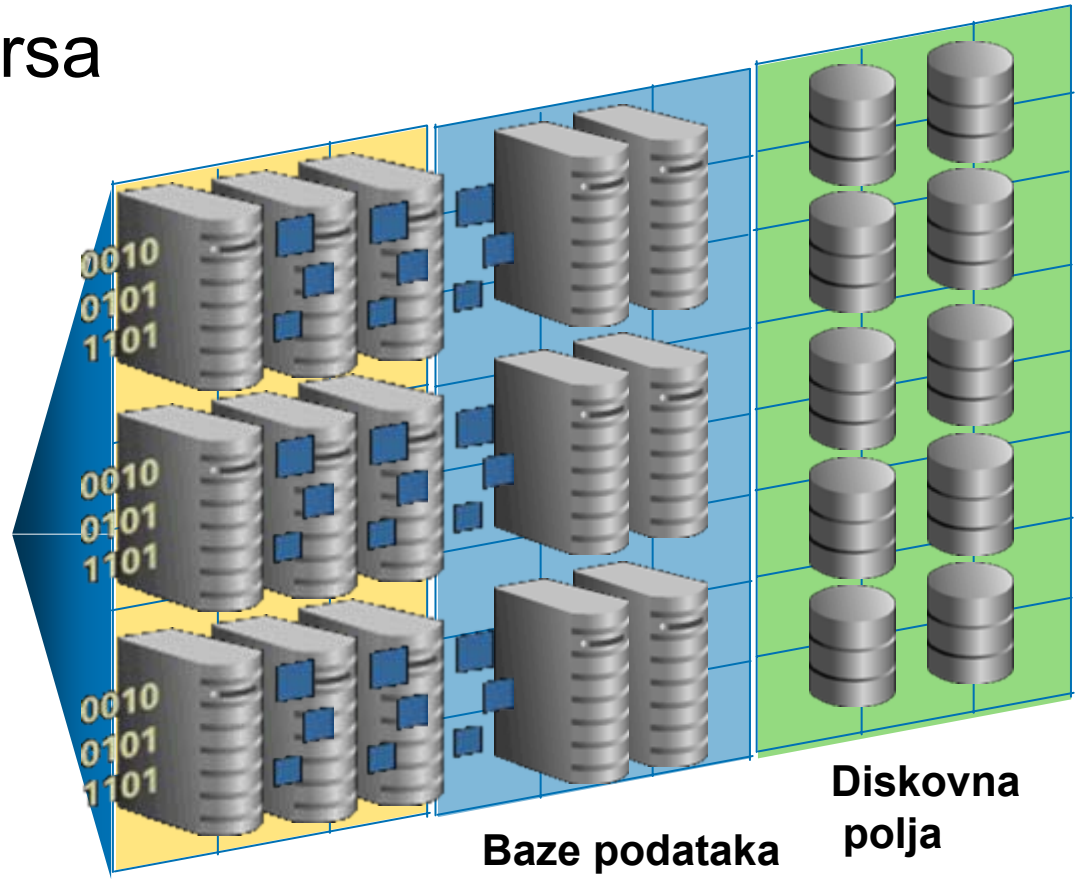


Oracle 11g grid svojstva

Oracle Grid Computing

- Virtualizacija
- Pooling/sharing resursa
- Dinamička dodjela resursa
- Automatizirano upravljanje

Grid Control

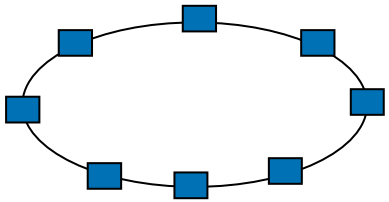


Aplikacijski poslužitelji

Baze podataka

Diskovna polja

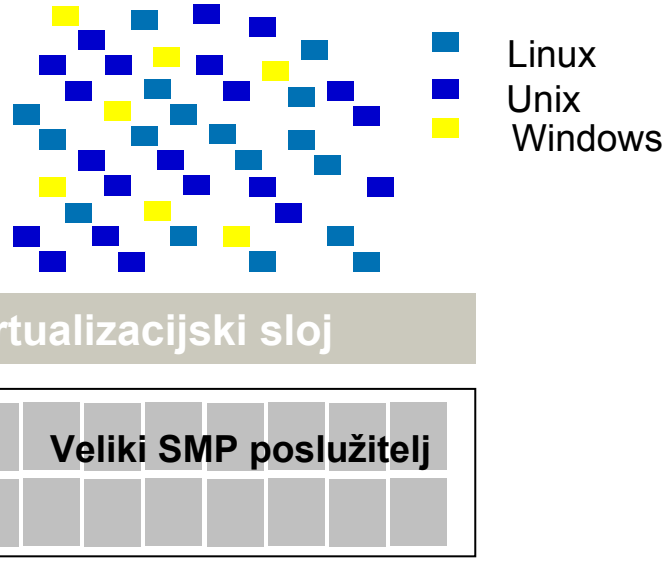
Grid virtualizacija



9i, 10g, 11g

Mnoštvo manjih poslužitelja se ponaša kao jedan veliki

Virtualizacija



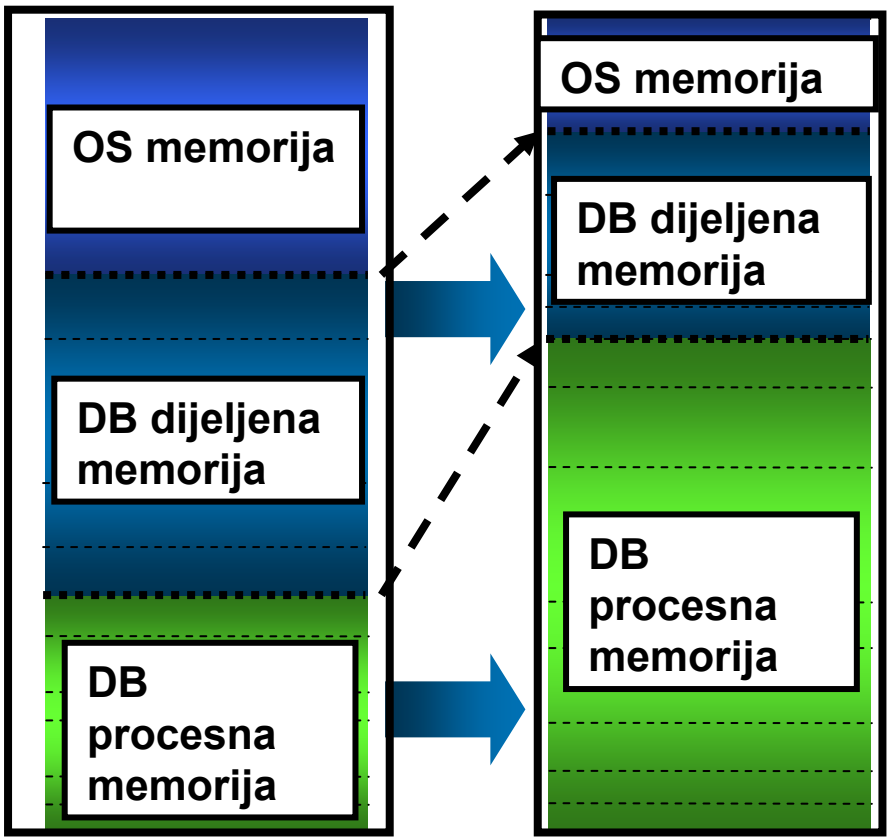
Jedan veliki poslužitelj se ponaša kao mnoštvo manjih

Virtualizacija preko više resursa

Virtualizacija unutar pojedinih resursa

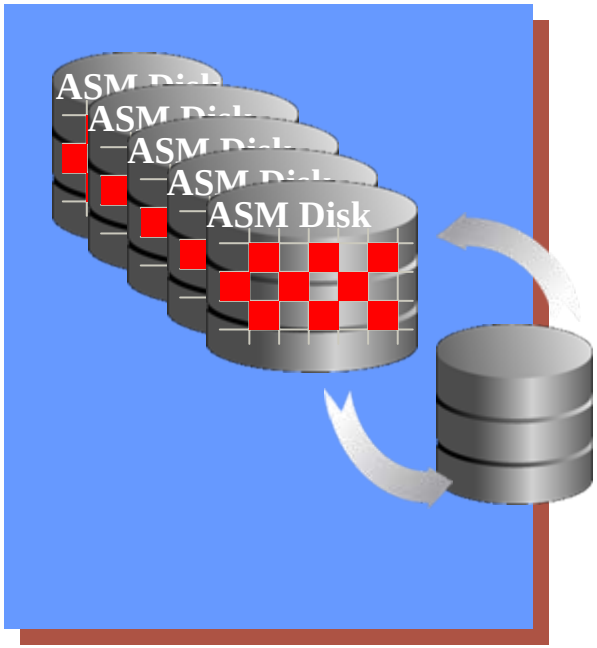
Oba pristupa se nadopunjuju, s ciljem boljeg iskorištenja resursa i učinkovitije konsolidacije.

Automatic Memory Tuning



- Automatska prilagodba promjenama opterećenja
- Učinkovitije korištenje memorije
- Prilagodba PGA, SGA, os memorije
- Single dynamic memory parameter

Automatic Storage Management



- Preferirana metoda
  - Lakše upravljanje od datotečnog sustava
  - Performanse raw volumena
  - Built-in
  - Dijeljeni storage pool za sve baze

  - >65% 10g RAC instalacija na ASM
  - >25% 10g korisnika upotrebljava ASM
  - Mnogo VLDB preko 10TB

Pregled HP c-class blade sustava

BladeSystem c-Class

c-Class šasija

HP BladeSystem c7000

Do 16 blade poslužitelja

Do 8 LAN/SAN preklopnika

BladeSystem c-Class šasija

HP BladeSystem c3000

Do 8 blade poslužitelja
Do 4 LAN/SAN preklopnika

Bolje performanse

Virtualizacija

- Više memorije
- Brži procesori
- Podrška za SD hypervisor

Bolje performanse

- Green computing
- Bolji TCO/ROI

Tehnologija

- Procesori nove generacije
- DDR3 memorija
- SSD (Solid State Drive)
- PCI-E Gen 2

Nehalem Processor

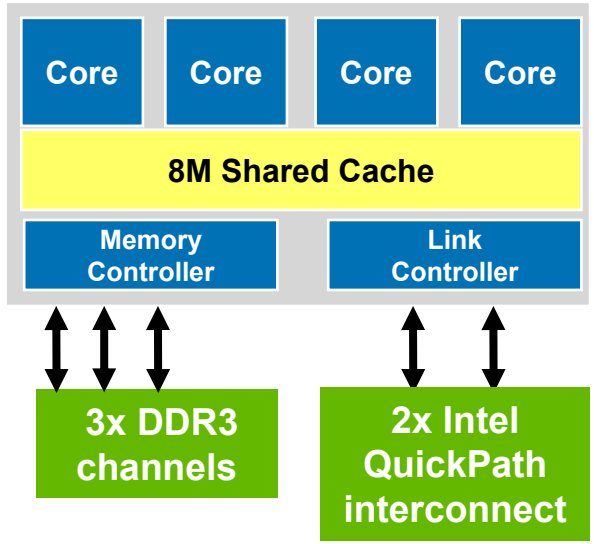
QuickPath arhitektura
Omogućava veću propusnost

Integrirani memorijski sklopovi
Odvojeni sklopovi za svaku jezgru

Turbo Mode
Privremeno povećanje takta po potrebi

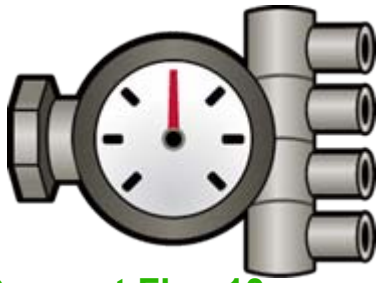
Dynamic Power Management
Omogućava turbo način rada

Hyper-Threading tehnologija
Povećane performanse kroz paralelizam



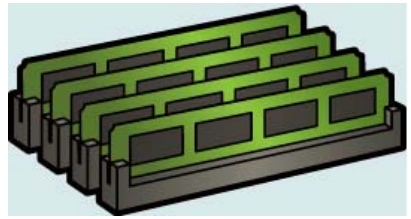
HP ProLiant BladeSystem G6

Novi koncepti



HP Virtual Connect Flex-10

Više memorije



50% to 100%

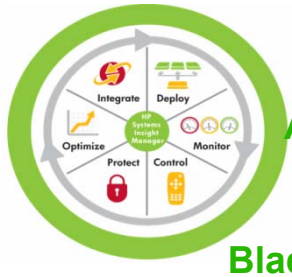
to DIMM slots

Bolje performanse, manja potrošnja

Next Gen Processors

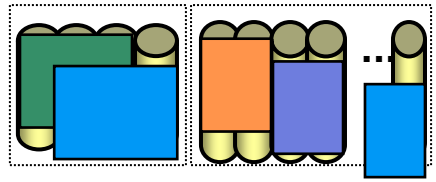
Dynamic Power Capping

Konsolidirano upravljanje



iLO Advanced for BladeSystem

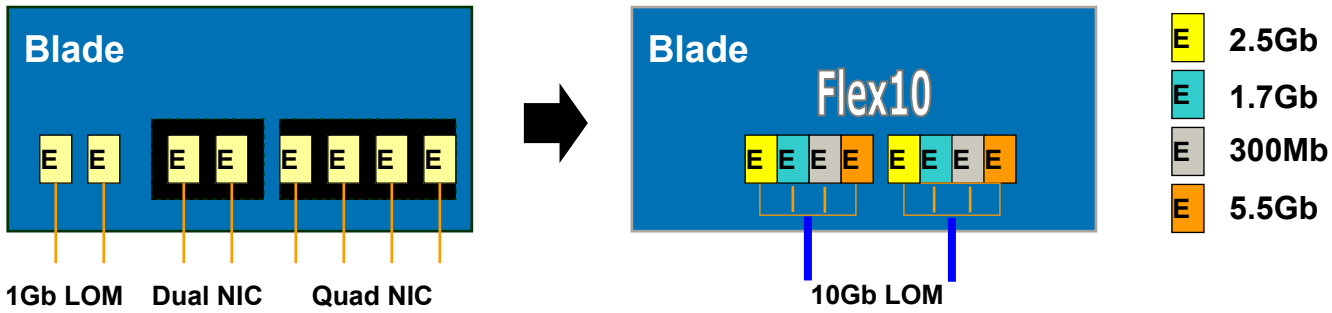
Novo SAS rješenje



HP BladeSystem shared SAS Storage

Prilagodljiva NIC (FlexNIC) tehnologija

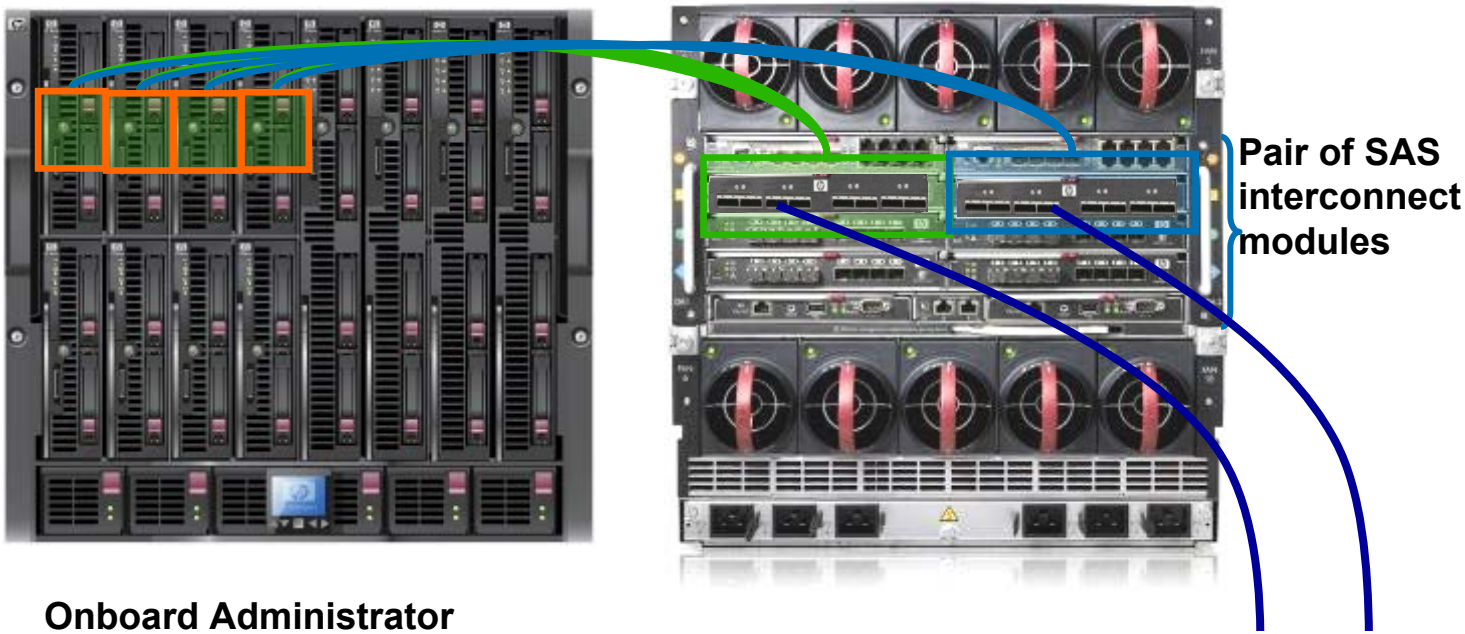
- 10Gb Ethernet
  - Potpuna implementacija PCI specifikacije
  - OS neovisan
  - Virtualni sklopovi podržani na BIOS razini
  - Brzine od 100Mb/s do 10Gb/s u koracima od 100Mb/s



SAS preklopnici

Blade Server

- **Each blade with SAS mezzanine card**



Onboard Administrator

- **Dijeljeno diskovno polje**
- **Svaki blade poslužitelj ima SAS HBA za spajanje na SAS preklopnike**
- **SAS preklopnici spojeni na MSA2000**

MSA2000sa SAS Array

Optimalna implementacija Oracle 11g RAC na HP blade sustavima

Testna konfiguracija

4 x BL480c RAC blade poslužitelja:

- 2 x 3.0Ghz dual-core Xeon, 16GB RAM, RHEL 4 Advanced Server

4 x BL460c blade poslužiteljas (Benchmark, SE, CV&SIM, OV):

- 2 x 3Ghz dual-core Xeon, 8GB RAM, Win2003

HP StorageWorks EVA8000:

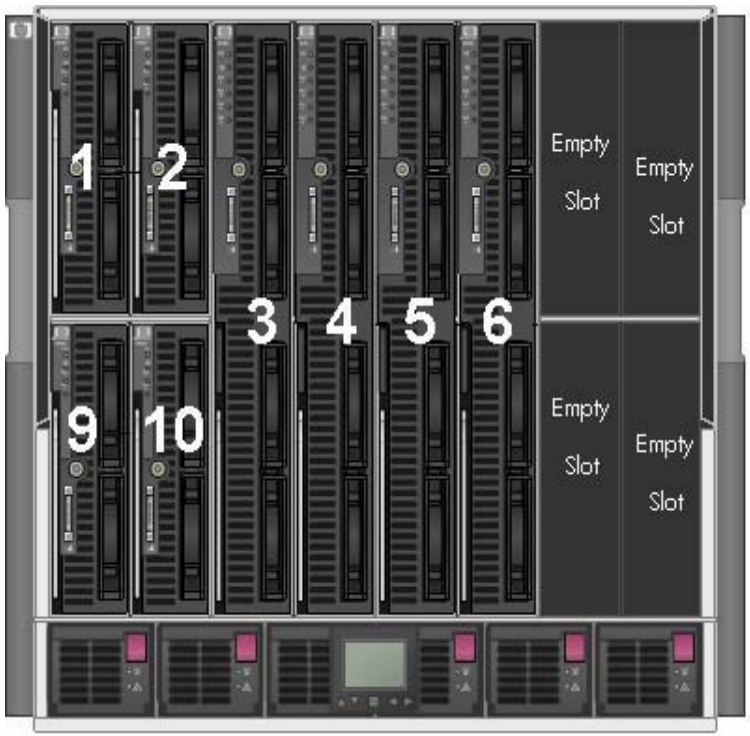
- DG1 (112 x 75GB 15K drives): user, system, data and redo log files

- DG2 (40 x 146GB 15K drives): archive files, flashback area i RMAN backup

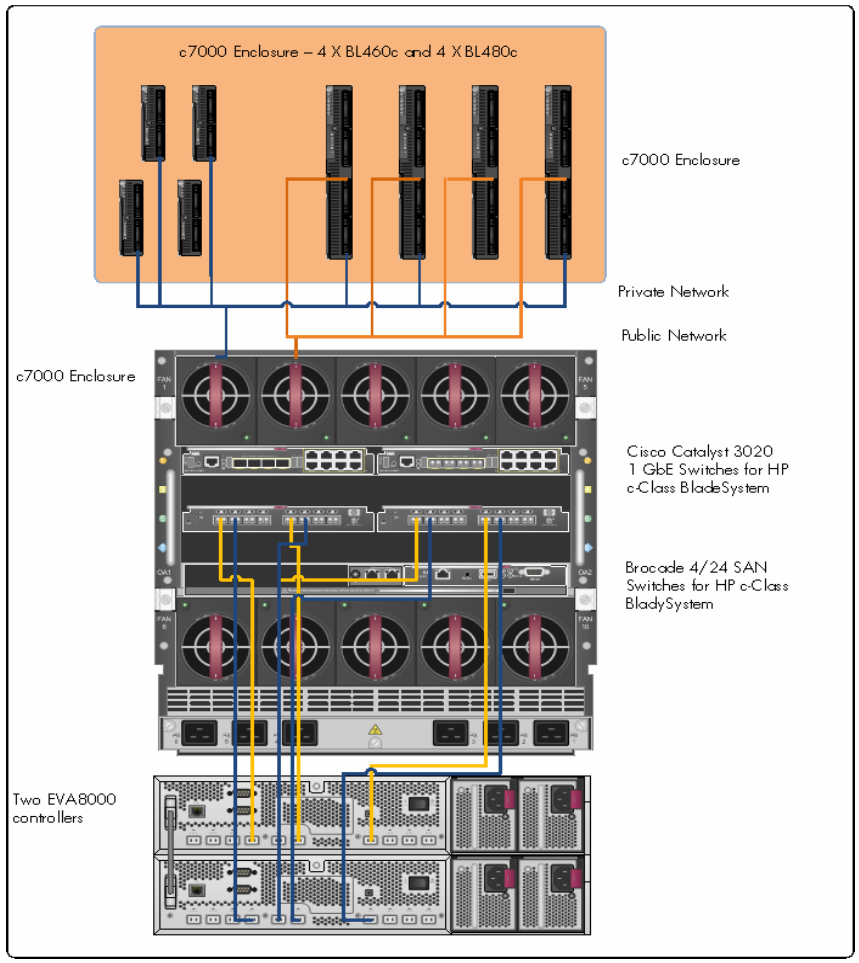
2 x 4Gb SAN preklopnici

Oracle 11g RAC

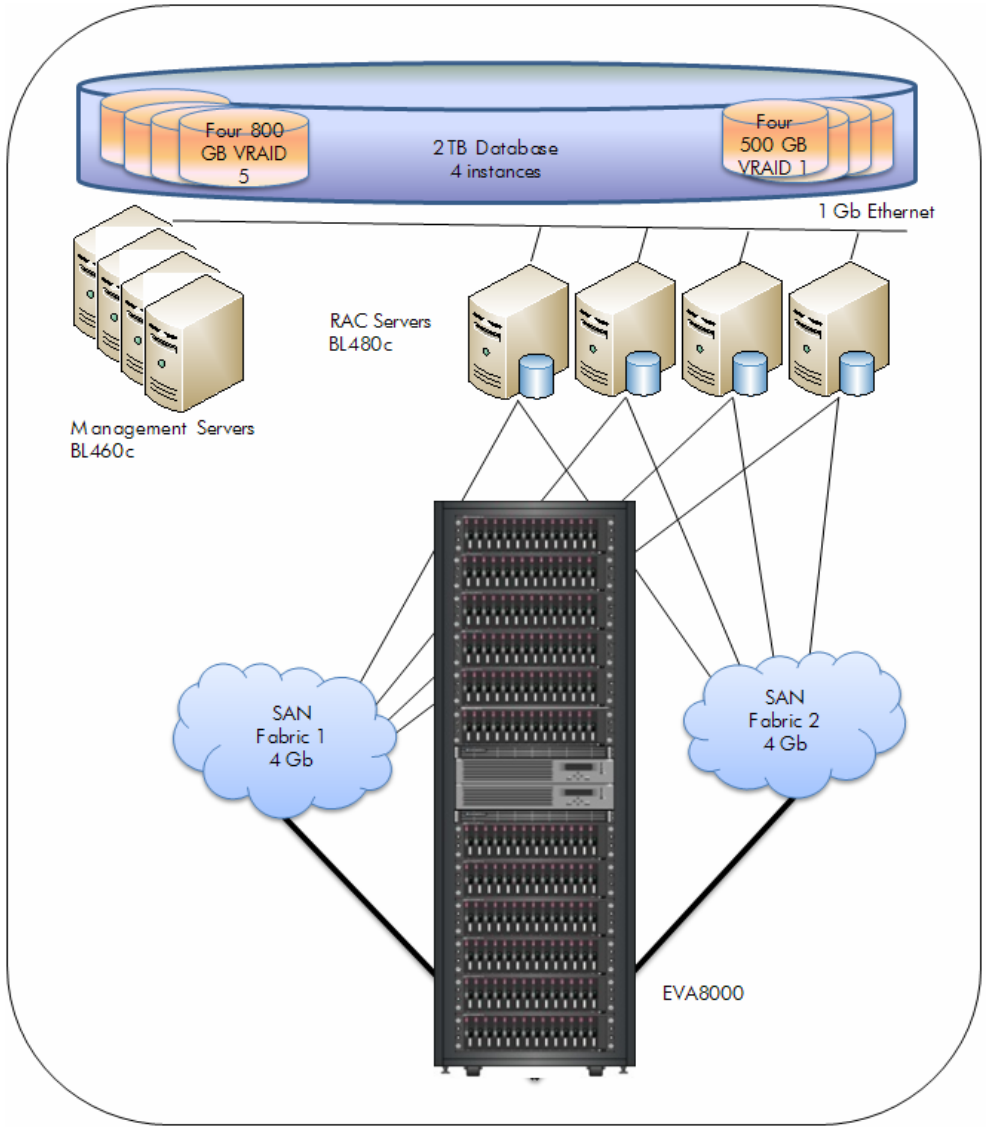
Prednji izgled c7000 šasije



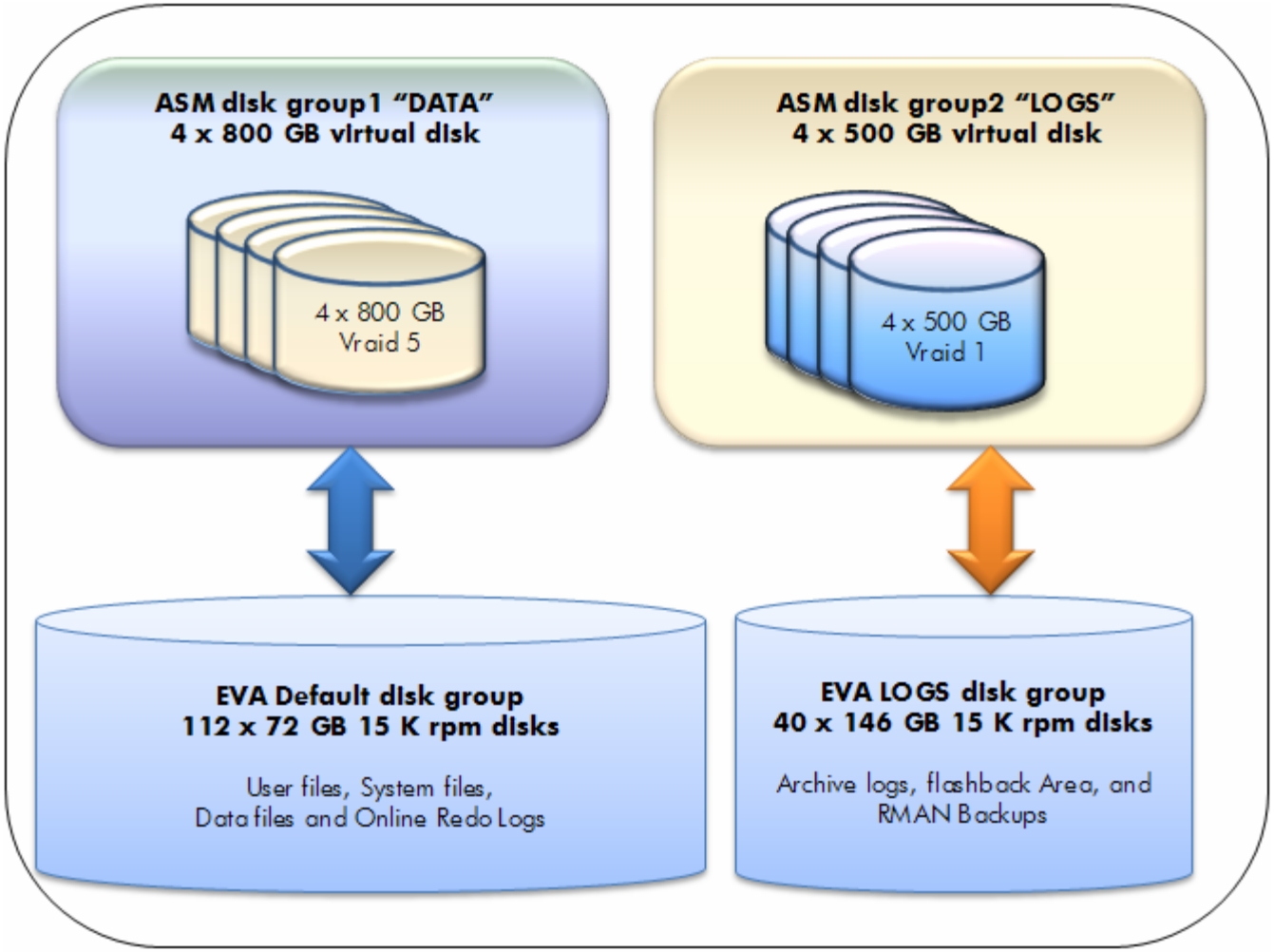
Shema spajanja



Logička shema testnog okoliša



Shema ASM i EVA disk grupa



Ciljevi i način testiranja

- Testiranje 11g RAC skalabilnosti na HP blade sustavu
- Povećanje opterećenja u koracima od po 200 korisnika, dok se ne postigne najveći #TPS
- Benchmarking softver: BMF od Questa

BMF postavke: 45.7% insert, 45.5% update, 8.8% read

Simulirano 100ms “think time” po transakciji

Metrike

- Oracle 11g RAC (BMF softver)
  - Transactions per second (TPS)
  - Transaction resonse time u [ms]

- 460c blade farma
  - Run queue length (Oracle Enterprise Managera)
  - CPU iskorištenje (HP Openview Performance Manager)
  - Memorijsko zauzeće (OEM & OPM)

- EVA8000
  - IOPS (EVAPerf)
  - Ukupna propusnost (EVAPerf)

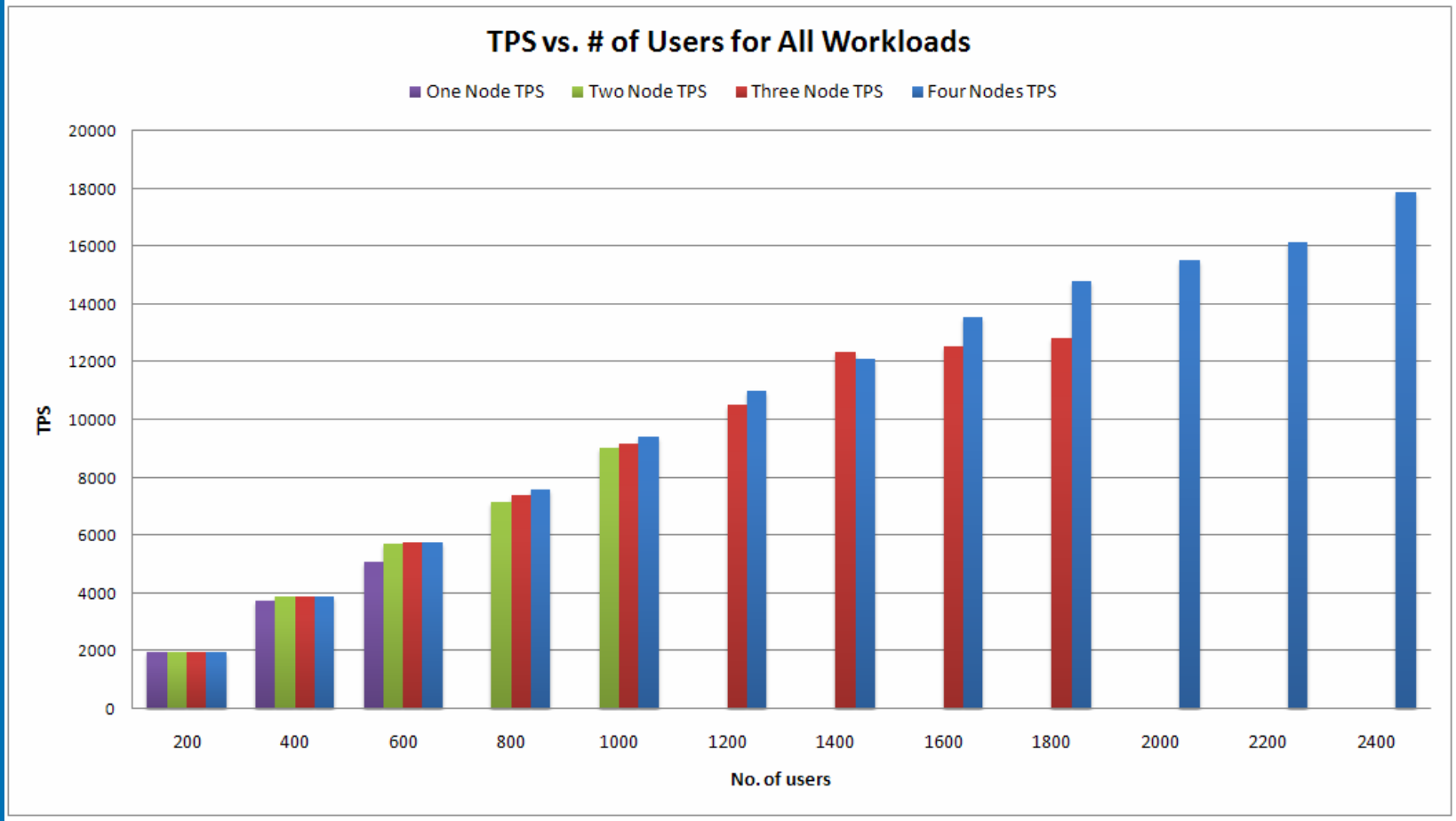
RAC parametri skaliranja

- Tri najvažnija parametra skaliranja za RAC:
  - #users
  - TPS
  - Transaction response time

Rezultati testa:

- #users skalira linearno sa brojem poslužitelja
- TPS: NE
- Transaction response time: NE

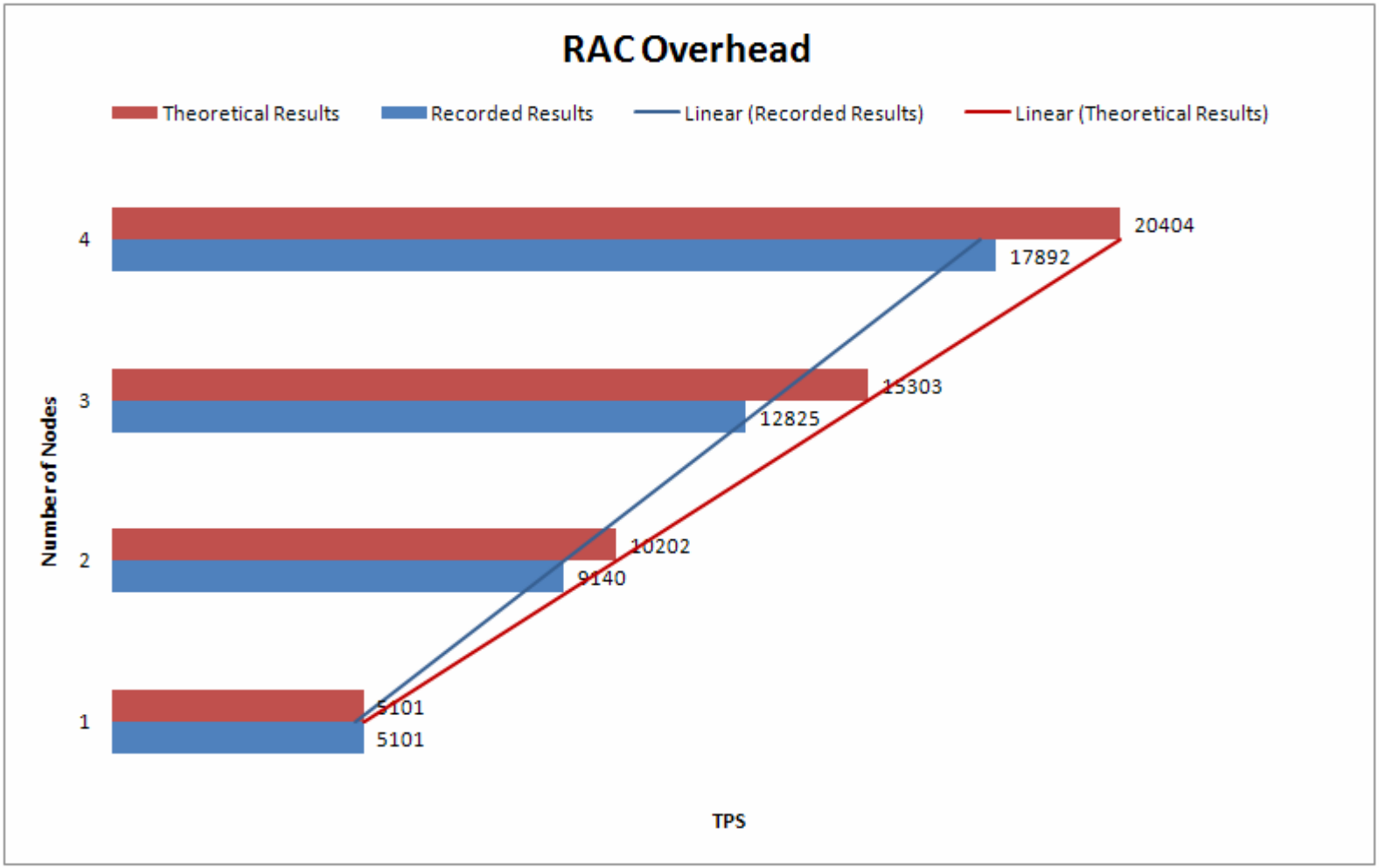
TPS vs. #users


TPS vs. # of Users for All Workloads
One Node TPS
Two Node TPS
Three Node TPS
Four Nodes TPS
TPS
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
No. of users

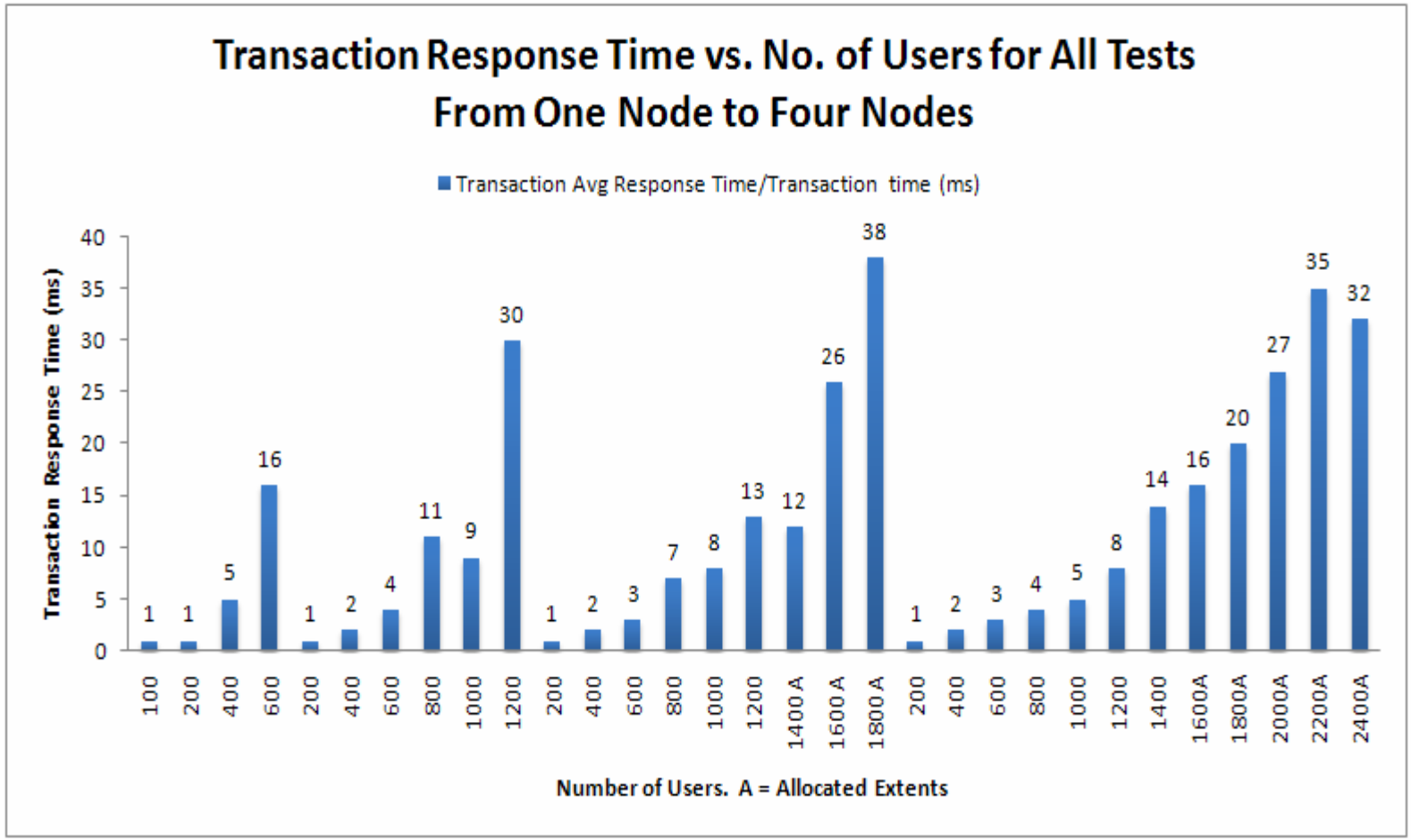
Skalabilnost & overhead

# 460c	max #users	Max #TPS za sve 460c	Max #TPS po 460c	Overhead (s obzirom na TPS)
1	600	5.101	5.101	Baseline 0%
2	1200	9.140	4.570	10.41%
3	1800	12.825	4.275	16.19%
4	2400	17.892	4.473	12.31%

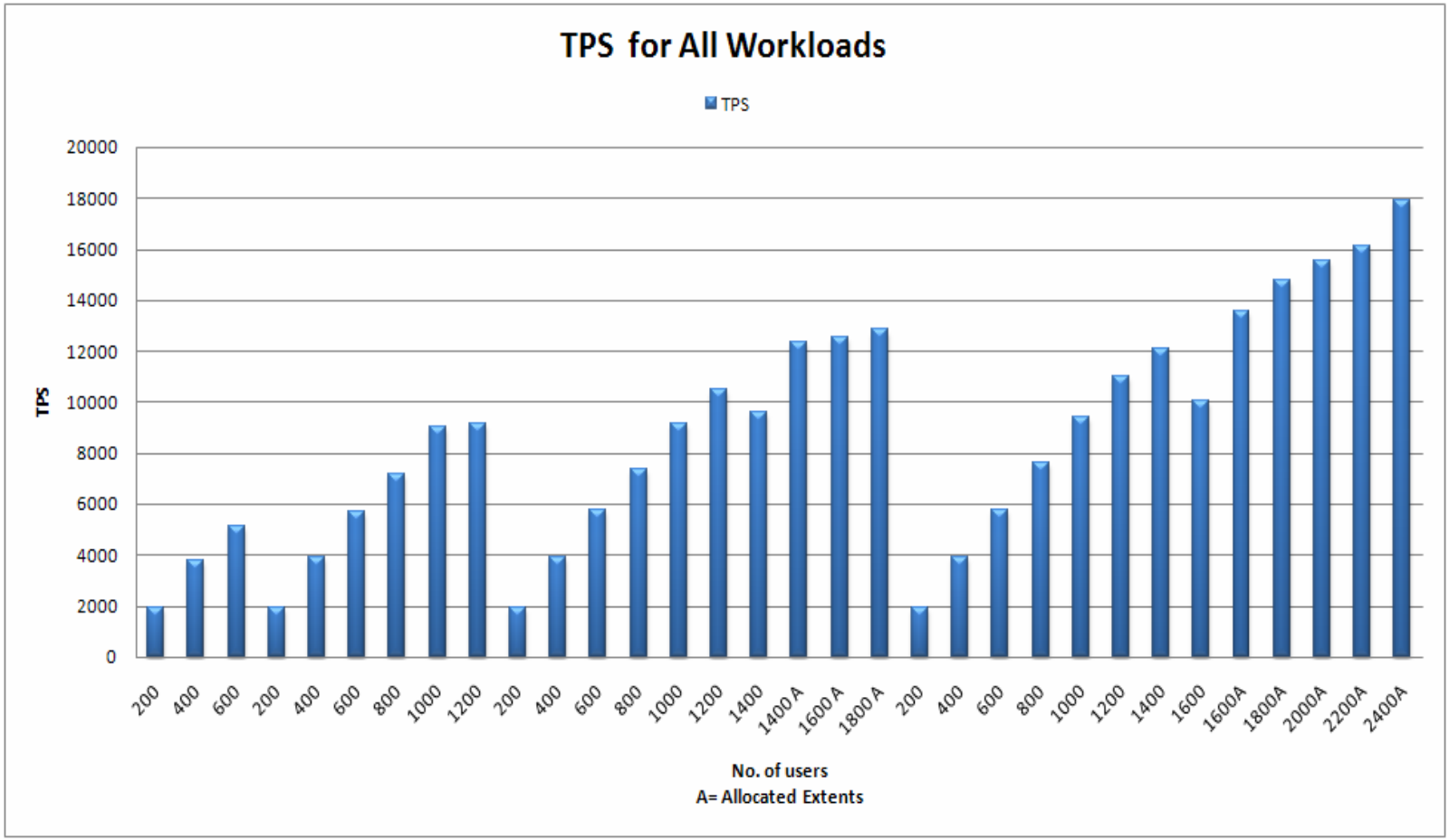
RAC overhead



Transaction response time

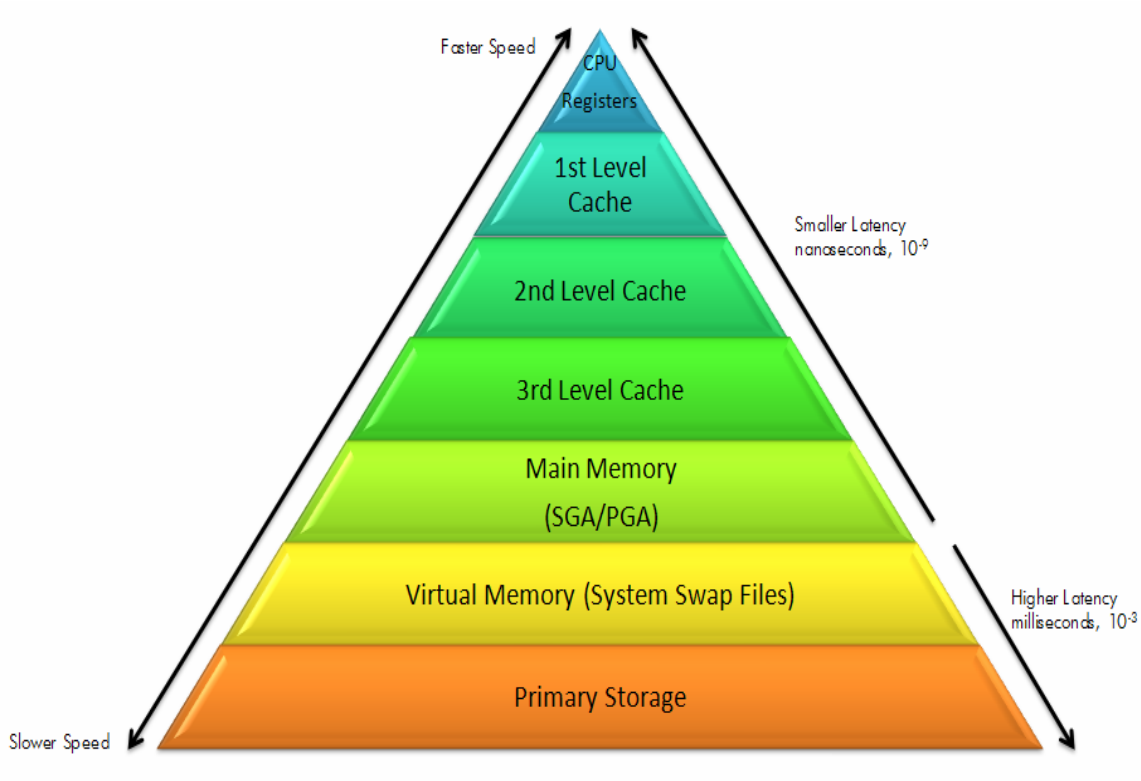


Safe TPS and safe #users

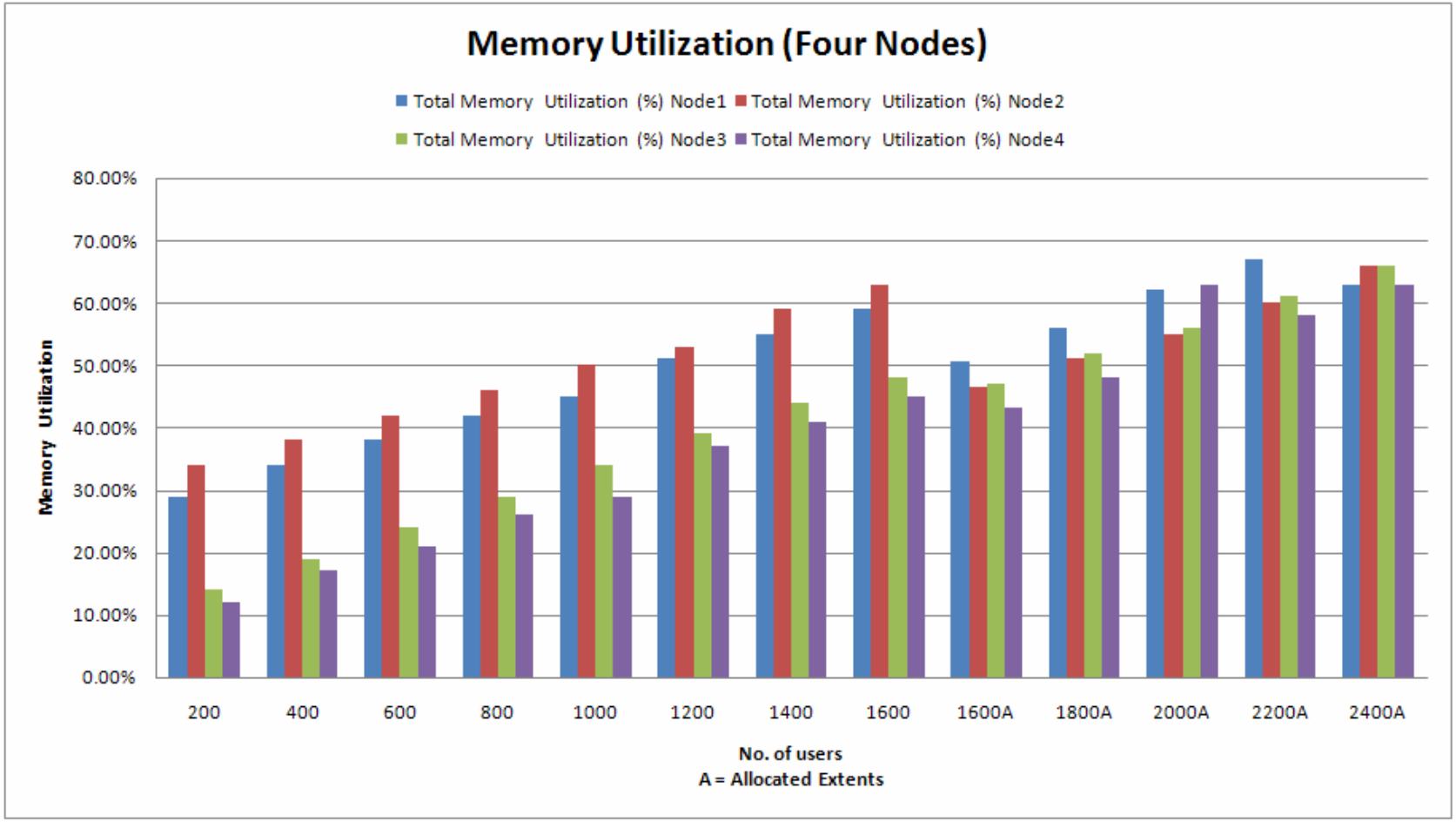


RAC: upravljanje memorijom

- Preporuka: Automatski balans između SGA i PGA
- Preporuka: smanjiti IPC (relativno spor, opterećuje procesore)

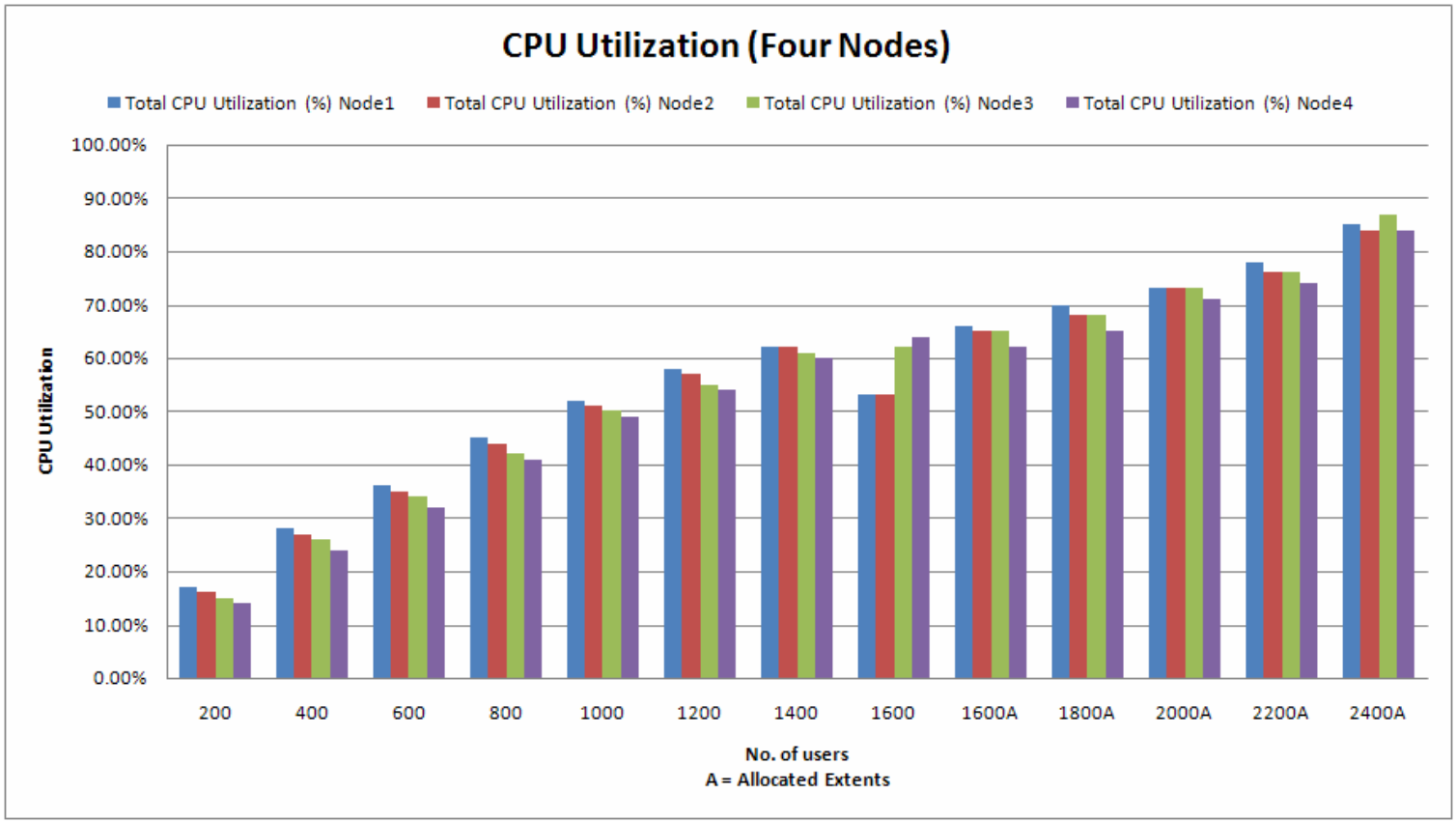


Zauzeće memorije u RAC klasteru



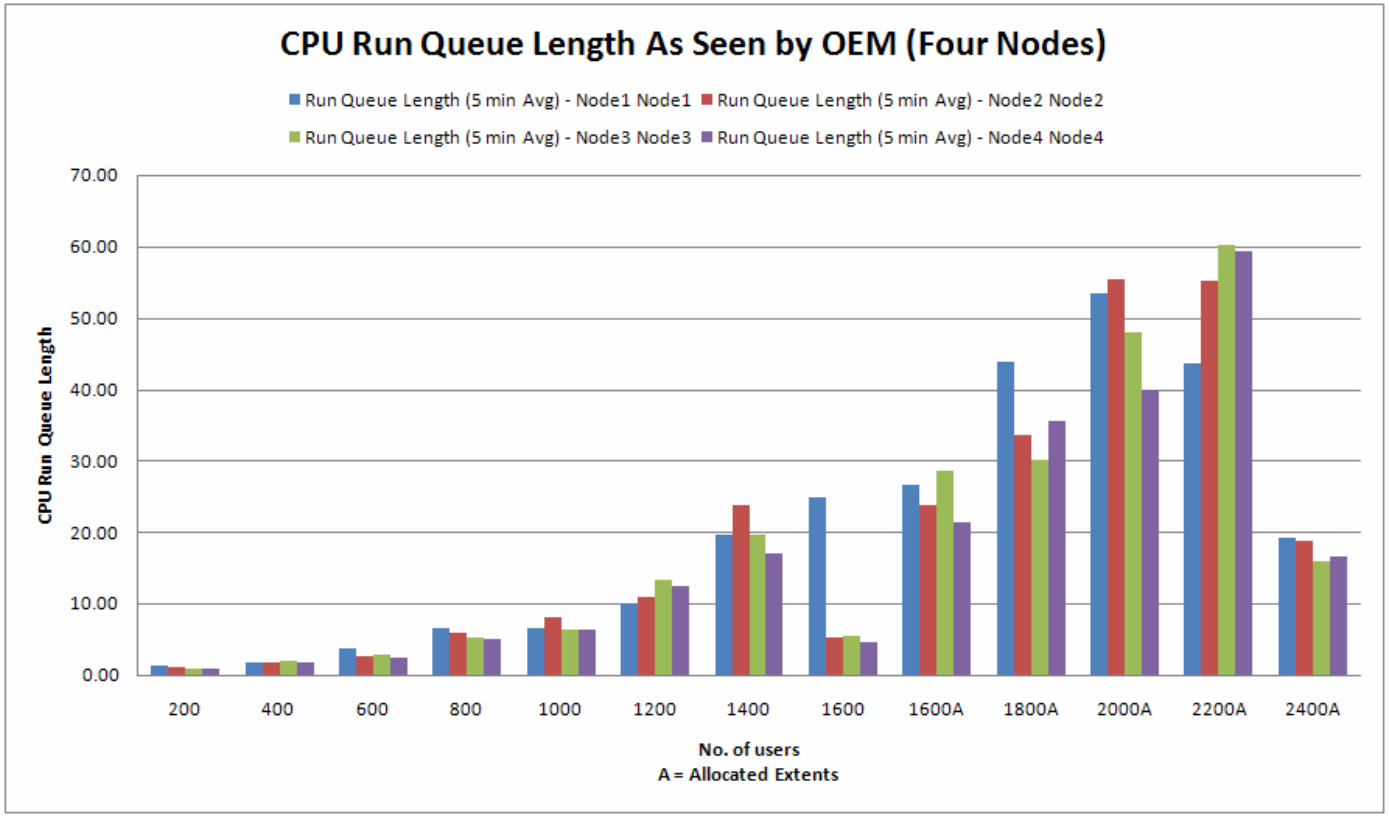
Neravnomjerno CPU opterećenje

- Ako razlika u CPU opterećenju među čvorovima prelazi 10%, dobro je istražiti vremena čekanja na bazi. Uobičajena razlika je oko 3% u OLTP okruženjima. Ovaj problem se može riješiti pre-alociranjem ekstenata.



CPU Run Queue Length razlike

- Ukazuje na preveliko vrijeme čekanja



Što smo naučili prilikom testiranja

- Prilikom dimenzioniranja RAC klaster resursa, dobro je započeti sa testiranjem mogućnosti jednog poslužitelja. Takav pristup omogućava dobru procjenu skalabilnosti RAC sustava: RAC klaster dimenzioniran na takav način može opsluživati N puta veći broj korisnika od jednog poslužitelja, sa maksimalnim dodatnim opterećenjem od 16%.

- Max. #TPS i #users treba odrediti prema zahtjevima front-end aplikacije. Dodavanjem drugog poslužitelja smanjuje se vrijeme odziva transakcije (transaction response time) za 57%. Dodavanjem trećeg i četvrtog poslužitelja to vrijeme se dodatno smanjuje za daljnjih 40% respektivno.

- Potrebno je poznavati mehanizam korištenja memorije od strane operacijskog sustava i front-end aplikacije. Različiti operacijski sustavi različito izvještavaju o veličini korištene memorije. Neke aplikacije daju izvještaj o korištenoj cache memoriji, druge ne. Treba osigurati dovoljno memorije za os i Oracle buffer cache. U suprotnom, swapping mehanizam izaziva neprihvatljivo povećanje read/write operacija prema relativno sporim diskovima. HP/Oracle RAC sizer predstavlja dobro ishodište za određenje potrebne memorije po blade poslužitelju.

- EVA diskovno polje i RAC klaster treba sinhronizirati sa istim davačem vremena (time server).

- Treba osigurati dovoljno memorije za os i Oracle buffer cache. U suprotnom, swapping mehanizam izaziva neprihvatljivo povećanje read/write operacija prema relativno sporim diskovima. HP/Oracle RAC sizer predstavlja dobro ishodište za određenje potrebne memorije po blade poslužitelju.

- Prihvatljiva razlika u CPU opterećenju među poslužiteljima je do 10% u OLTP okolišu. Ukoliko je ona veća, potrebno je ispitati vremena čekanja u bazi. <CPU Run Queue Length> razlike također indiciraju ovaj problem. Može pomoći pre-alociranje ekstenata.

Demo