Modelarea şi performanțele sistemelor de stocare

Ceea ce urmărim prin prezentul material este să oferim pe cât posibil un ghid de modelare şi selecție a sistemelor de stocare funcție de necesitățile aplicațiilor folosite, beneficiind de înțelegerea teoriei cozilor explicată în prima parte a materialului. În final se va demonstra o analiză comparativă, prin teste, între două tipuri de sisteme de stocare, rotaționale în configurație RAID şi SSD.

Teoria cozilor şi utilizarea resurselor

Chiar dacă teoria cozilor poate fi uneori prea academică şi prea matematică, ea include un număr de reguli de bază care descriu comportarea sistemelor şi a componentelor acestora la niveluri ridicate de utilizare.

Una dîntre principalele concluzii ale teoriei cozilor este că nu se poate ajunge la un nivel mare de utilizare a unei resurse fără ca un anumit fel de coadă de aşteptare să apară. Poate că nu pare evident însă acesta este un adevăr. În esență o coadă de cereri în aşteptare este necesară pentru a ține resursa ocupată, în aşa fel încît să existe întotdeauna ceva de făcut după ce se termină activitatea curentă.

Cu cât utilizarea unei resurse (procesor, sistem de stocare, memorie, rețea) se apropie de 100% cu atât lucrurile devin mai îngrijorătoare pentru că asta înseamnă în general un număr foarte mare de cereri care aşteaptă în diverse cozi.

Efectul direct al unor cozi lungi este aşteptarea din ce în ce mai îndelungată pentru rezolvarea unei cereri de calcul. Timpul total de răspuns pentru o cerere este acum egal cu durata de aşteptare în coadă adaugată la durata normală de procesare. Dacă intr-o coadă sunt în mod constant cite 5 cereri spre exemplu, timpul de răspuns pentru fiecare cerere este în fapt de 5 ori mai lung decât durata normală de procesare a unei cereri de către sistem.

Teoria cozilor ne arată formula pentru numărul total de cereri (în coadă şi în curs de procesare – **N**) relativ la gradul de utilizare al unei resurse (**U**):

N = U / (1 - U)

Formula pentru timpul total de răspuns (R) funcție de gradul de utilizare şi de timpul efectiv de prelucrare (**T**) este:

R = T / (1 - U)

Prin urmare un grad mare de utilizare conduce la cozi lungi şi la timpi mari de răspuns.

Din cele de mai sus rezultă că scopul urmărit în momentul modelării unui sistem de calcul nu este timpul de tranzacționare (**T**) ci timpul total de răspuns pentru o tranzacție (**R**).

Teoria cozilor vine prin urmare să confirme faptul că atunci când un sistem este scalat pentru a procesa mai multe date intotdeauna va apare o gâtuire la o anumită componentă care va fi utilizată aproape de capacitatea sa maximă, componenta la care se vor forma cozi şi timpul de răspuns general va creşte până la valori inacceptabile. Sarcina noastră va fi atunci să identificăm şi să remediem aceste gâtuiri.

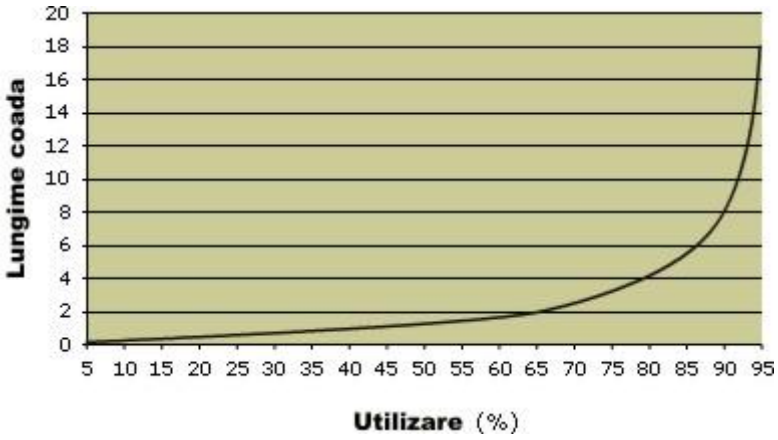


Figura 1

Lungimea exponențială a cozilor funcție de gradul de utilizare a unei resurse

Niveluri de utilizare a procesoarelor

Conform figurii 1 este recomandat ca utilizarea procesoarelor să fie mereu menținută sub 75%. Este adevărat ca ocazional pot apărea vârfuri de utilizare peste 75% dar este important ca ele să reprezinte numai nişte excepții. Deoarece la 75% utilizare coada de aşteptare are deja 3 cereri, înseamnă că procesorul rezolvă orice sarcină de 3 ori mai încet decât la niveluri scăzute de utilizare, iar peste 75% viteza scade în mod exponențial.

Niveluri de utilizare a discurilor

Nivelul de utilizare al discurilor devine problematic în mod exponențial dupa 85%. Acest lucru este adevărat atât pentru gradul de încărcare a spațiului pe disc cât şi pentru gradul de încărcare al interfeței de comunicație.

Niveluri de utilizare a memoriei

Erorile de paginare (*page faults*) apar atunci când nu se pot găsi datele în memorie şi ele trebuiesc recuperate de pe disc. Cum datele din memorie pot fi accesate cu viteze de 1000 de ori mai mari decât de pe disc se doreşte, în consecință, creşterea volumului de memorie a computerelor pentru a evita cît mai mult apariția acestor erori. Se recomandă menținerea unui nivel de încărcare sub 75%.

Sisteme de stocare, parametri şi corelații

Sistemele de stocare au cunoscut o dezvoltare spectaculoasă în ultimii ani, folosind tehnologii din cele mai variate cum ar fi suporturile magnetice, optice dar şi stocarea în memorii flash sau RAM cunoscute în general ca SSD.

Indiferent de tehnologia în care sunt construite sistemele de stocare, ele vor prezenta mereu citeva caracteristici de bază relativ la două componente de importanță majoră:

- Lanțul de interfețe de conectare a mediului de stocare la computer
- Mediul de stocare propriu-zis (magnetic, optic, cip de memorie)

Ori de câte ori vom studia o fişă tehnică a unui sistem de stocare vom descoperi diverşi parametri legați de cele două subsisteme menționate mai sus.

Se constată de cele mai multe ori că parametrii respectivi sunt extrem de variați şi că suntem cel mai adesea puşi în dificultate în luarea unei decizii corecte relativ la modelarea unui sistem de stocare care să ofere performanța necesară la un preț acceptabil.

În practică se întilnesc citeva tipuri de aplicații care au cerinte destul de diferite de la un sistem de stocare:

- Aplicații tranzacționale
- Aplicații de raportare
- Aplicații media (streaming, transferuri de fişiere mari)
- Aplicații de tip bibliotecă sau aplicații de arhivare

Desigur că natura aplicației va dicta care este cel mai important parametru al sistemului de stocare pe care urmărim să îl modelăm. În figura 2 sunt reprezentate gradul de distribuție al acestor tipuri de aplicații precum şi nevoia lor pentru sisteme de stocare performante.

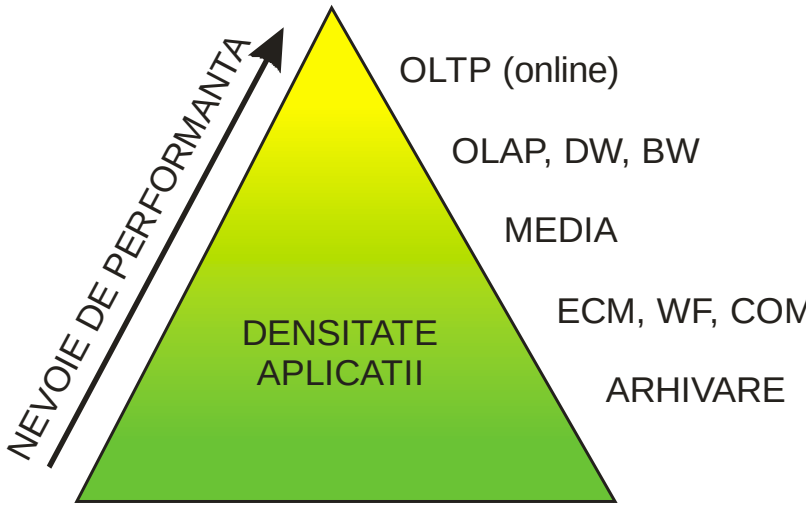


Figura 2

Tipuri de aplicații şi nevoi de performanță din partea sistemelor de stocare

Dacă ne limităm totuşi la o singură aplicație pe care o considerăm critică şi care foloseşte un volum extrem de mare de date putem să facem o analiză a fişierelor utilizate de către aplicație din punctul de vedere al frecvenței de acces. După o astfel de analiză se poate modela cu succes un mediu de stocare format din echipamente de tipuri şi cu performanțe diferite care să sustină aplicația cu o performanță totală rezonabilă şi cu un cost rezonabil (vezi figura 3).

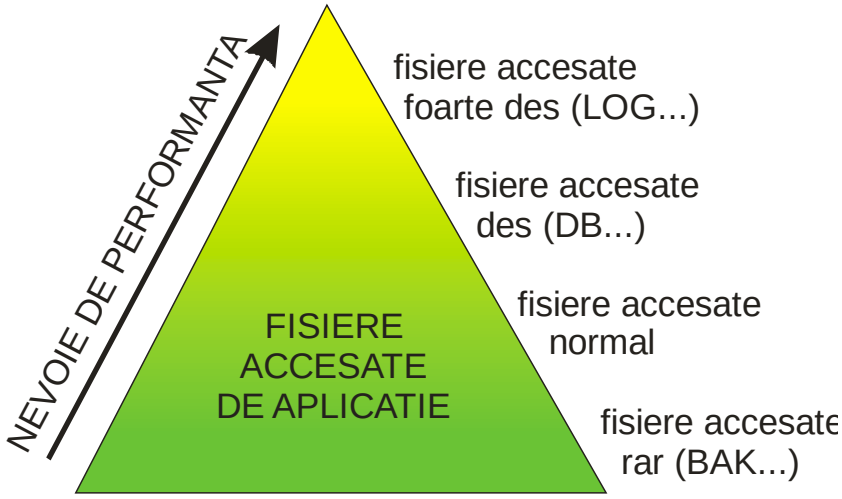


Figura 3

Distribuția frecvenței de acces la fişiere pentru o aplicație critică

Aplicații tranzacționale (OLTP)

În acest sens, pentru o aplicație tranzacțională va fi în primul rând necesar ca timpul de răspuns (**R)** să fie cât mai mic. Acest lucru se poate obține folosind un sistem de stocare cu o latența (**T**) cât mai mică şi cu un grad de utilizare cât mai mic. Cum din considerente economice nu ne permitem să menținem echipamentele neutilizate, rezultă că parametrul cel mai important al sistemului de stocare pentru o aplicație tranzacțională este latența (**T**).

Latența este un parametru care depinde de însumarea timpilor de acces la mediul de stocare şi a timpilor de transfer dinspre mediul de stocare înspre interfața de conectare. Prin urmare interfața de conectare la computer trebuie în primul rind să aibă latența mică şi nu în mod neapărat o bandă foarte mare, deoarece în sistemele tranzacționale mărimea tranzacțiilor este redusă.

Dacă studiem fişele tehnice de la majoritatea sistemelor de stocare (SAN) constatăm că acest parametru este în general trecut cu vederea şi că toți producatorii se întrec să prezinte un alt parametru care exprimă capabilitatea de a susține un număr de tranzacții în unitatea de timp, notat **IOPS** (Input Output Per Second).

Acest parametru este conectat de latență prin legea lui Little: **Q = T x IOPS**, unde **Q** este numărul de cereri care se află în coadă la un moment dat (reamintim că **N = Q + numărul de cereri aflate deja în prelucrare**).

Atragem atenția aici că **IOPS** este invers proporțional cu latența **T** şi că pentru un sistem cu latența mare, pentru a obține **IOPS** mare trebuie în general construită o configurație uriaşă, care de cele mai multe ori este doar pur teoretică şi nu va fi niciodată implementată datorită costurilor enorme.

În concluzie trebuie urmărit mai curind parametrul latență şi nu parametrul **IOPS** dacă dorim să modelăm un sistem de stocare pentru o aplicație tranzacțională.

Un alt aspect important pentru o aplicație tranzacțională este proporția comenzilor de scriere față de cele de citire. Deoarece anumite medii de stocare prezintă latențe diferite la scriere şi la citire, este important, în anumite cazuri, de evaluat în mod distinct impactul timpilor la scriere şi la citire.

Nu în ultimul rând, aplicațiile tranzacționale necesită de regulă volume de stocare mai mici.

Dacă în mod excepțional aplicația are nevoie să tranzacționeze volume de date mari (de exemplu la baza de date nu se foloseşte o mărime de bloc uzuală de 4K sau 8K ci de 64K sau chiar mai mare) atunci trebuie luat în considerare şi parametrul **MBPS** care trebuie estimat după formula aproximativă:

MBPS = [DB block size] x IOPS

Aplicații de raportare (OLAP, BW, DW)

Aplicațiile de raportare seamănă cu cele tranzacționale cu excepția faptului că în majoritatea timpului au loc tranzacții de citire.

O altă diferență constă în faptul că volumul de stocare necesar este mult mai mare, prin urmare va trebui ales, din considerente economice, un mediu de stocare mai ieftin chiar dacă asta presupune un compromis în ceea ce priveşte parametrul latență (**T**).

Aplicațiile media

Cele mai mari consumatoare de bandă de transfer sunt aplicațiile media deoarece ele „tranzacționeaza” rar dar cu fişiere de mari dimensiuni.

Din punctul de vedere al unui sistem de stocare cel mai important parametru aici este lațimea de bandă oferită de interfața de conectare a sistemului la computer. În general găsim în fişele tehnice ale produselor de stocare acest parametru notat **MBPS** (MB Per Second).

Aplicații de bibliotecă şi de arhivare (ECM, WF, EMAIL, etc)

În ceea ce priveşte aplicațiile de arhivare, cel mai important parametru este prețul cât mai mic per MB şi stabilitatea cât mai mare a datelor pe termen lung. Aici de obicei părăsim lumea discurilor magnetice pentru a discuta despre benzi şi discuri optice.

Tipuri de sisteme de stocare

Sistemele de stocare folosite curent în mediul corporatist sunt denumite SAN şi oferă resurse de stocare pentru mai multe servere simultan.

Conectivitatea dinspre computer spre SAN este de cele mai multe ori oferită prin interfețe de mare viteză, denumite Fibre Channel (FC) sau Infiniband.

Chiar dacă la prima vedere două SAN-uri cu acelaşi număr de interfețe FC de 4Gbps spre exemplu par să fie identice din punct de vedere al conectivitătii, trebuie totuşi să oferim puțină atenție mai multor componente de conectivitate ale sistemului, unele chiar necesitând consultarea unor documente tehnice mai puțin vizibile la producători:

- Backplane
- Controllere aflate între backplane şi mediile de stocare
- Interfețele mediilor de stocare propriu-zise

De exemplu este inutil să populăm un SAN cu mai multe interfețe externe să spunem FC a căror viteză de transfer cumulată depăşeşte viteza de transfer a backplane-ului fără să avem şi alte considerente dincolo de cele de performanță.

Funcție de mediile de stocare folosite, SAN-urile pot fi grupate în:

- SAN bazat pe un număr de discuri fixe în tehnologie SCSI, SAS, etc
- SAN compatibil cu discurile fixe dar care foloseşte discuri flash SSD în loc de discurile fixe
- SAN SSD cu memorie de tip FLASH integrată proprietar în sistem
- SAN SSD cu memorie de tip DDR integrată proprietar în sistem

Putem bănui de la prima vedere că atât performanțele cât şi prețul per GB creşte în ordinea listei de mai sus. Cum se poate face un compromis rezonabil preț/performanță?

În figura 4 este prezentată o piramidă a performanțelor care poate fi suprapusă peste piramida tipurilor de Aplicații sau a densității de acces la fişiere pentru o aplicație critică. În acest fel datele cele mai accesate pot fi stocate pe un SAN foarte rapid (de exemplu RAM SSD) în timp ce datele cele mai puțin folosite pot fi stocate pe sisteme rotaționale clasice.

Modelarea unei structuri piramidale cu toate etajele de performanța poate deveni un factor determinant în a oferi un raport preț/calitate foarte bun tuturor tipurilor de Aplicații rulate într-un centru de calcul.

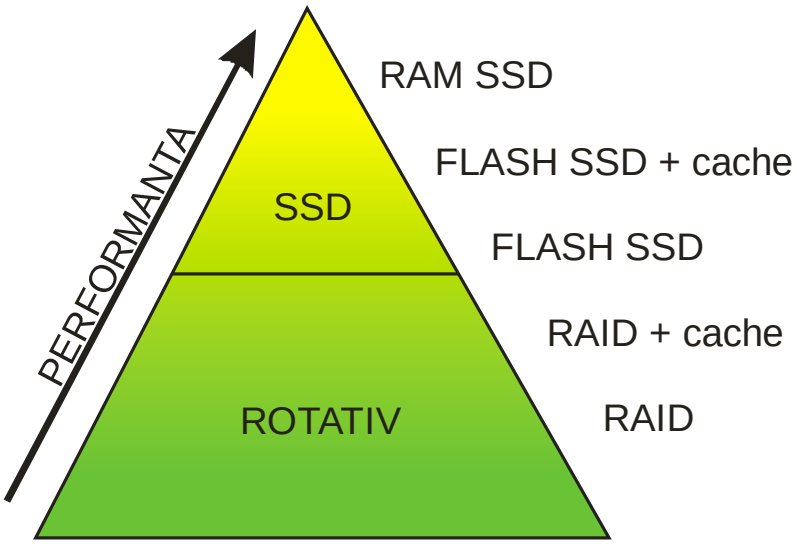


Figura 4

Performanța sistemelor de stocare funcție de tehnologia de stocare

Testarea sistemelor de stocare

Deşi fişele tehnice şi ofertele comerciale ne pot face o anumită idee despre performanțele unui sistem de stocare, este întotdeauna interesant de testat un sistem de stocare nou pentru a putea evalua impactul concret al noilor performanțe în propriul mediu de lucru.

Pentru acest lucru ne putem folosi de simplele instrumente de „benchmarking" care se găsesc pe internet dar care în general nu ne lasa sa intelegem prea usor ce se petrece cu sistemul respectiv şi din ce cauza apar limitari sau dacă exista solutii de imbunatatire a configuratiei pentru a ne atinge scopul propus.

O alta abordare este sa folosim instrumente de testare bazate pe teoria cozilor. Principiul de funcționare al acestora este aplicarea unui stres pe sistemul de stocare prin crearea şi menținerea constantă a unei cozi de cereri de tranzacționare (scriere şi citire).

Un avantaj al acestor unelte este acela că prin mărirea progresivă a dimensiunii cozilor, putem să depăşim la un moment dat avantajul oferit de componentele de caching şi să evaluăm comportarea sistemelor de stocare în cele mai rele condiții de lucru.

Când ne propunem să simulăm mediul de lucru existent, va trebui să răspundem la anumite întrebări:

- Care este proporția de cereri de scriere din totalul tranzacțiilor I/O?
- Care este dimensiunea medie a tranzacțiilor şi care este dispersia acestei dimensiuni?
- Care este numărul de cereri din coadă pentru care sistemul încă răspunde la un timp apropiat de latența (**T**) – aici de exemplu pentru un RAID 5 compus din 5 discuri, fiecare având latența 4ms, numărul optim de cereri în coadă va fi 4, deoarece logica ariei de discuri va distribui cele 4 cereri în mod echitabil pe discurile fizice, tot sistemul răspunzând la un timp apropiat de 4ms.

După ce evaluăm răspunsurile trebuie să construim un model de testare cu cereri mixte şi cu cozi de lungimi diferite pentru a acoperi cit mult situația aflată în mediul de lucru actual.

În continuare sunt prezentate grafice rezultate în urma testării comparative a două sisteme de stocare cu urmatorii parametrii:

Configurarea testului:

- Baza de date cu dimensiunea de 120GB
- Simulare aplicație de raportare (100% citiri)
- Mix de tranzacții mici (8K – 1...60 tranzacții) şi tranzacții mari (1M – 1...12 tranzacții)
- Windows Server 2008
- Fişierul swap nu se află pe nici unul dîntre cele două discuri

Sistemele testate:

- RAID 5, format din 5 discuri SAS 146GB, 15000rpm, controller cu 500MB cache
- Placa PCIe x4 cu memorii FLASH oferind un volum cu o capacitate de 450GB

Valorile de catalog ale parametrilor de bază pentru echipamentele testate:

Parametru (max)	HDD	RAID 5 (estimat)	FLASH
Latența	2 ms latența seq 3.4 ms seek time	2ms latența seq 3.4ms seek time	50 us latența random No seek time
IOPS	?	?	120 000 / 4KB citire 50000 / 4KB scriere
MBPS	SAS 3.0 Gb/s = 333 MBPS (8+1 bits)	?	700 MBPS

În graficele de mai jos, pe axa orizontală este exprimat numărul de tranzacții mici (8K) din coadă în timp ce menținem numărul de tranzacții mari (1M) din coadă constant, acesta fiind exprimat de titlul curbei.

Spre exemplu în graficul de latența FLASH, curba roşie 1 reprezintă timpul de latență constatat atunci când menținem în coadă 1 tranzacție mare de 1MB şi un număr variabil între 1 şi 60 de tranzacții mici de 8KB. Comparativ, acelaşi lucru este reprezentat în graficul de latență RAID.

De remarcat la graficul de latență RAID este ca o coadă de 4 cereri mici şi este optimă pentru sistemul RAID datorită logicii de paralelizare a efortului pe toate discurile.

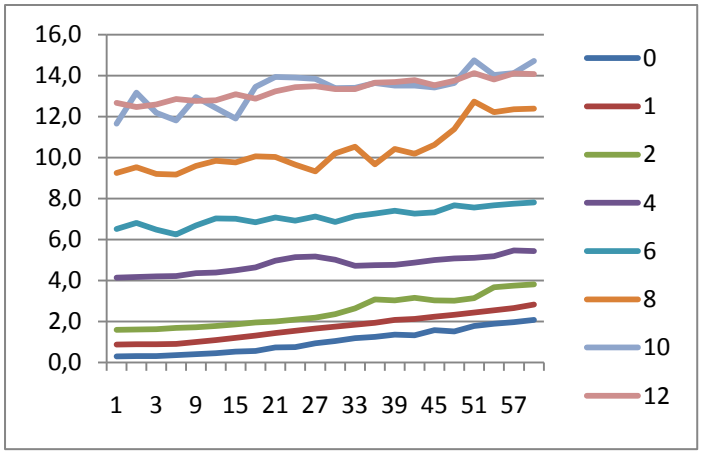
Un alt aspect interesant apare în a decide dacă plafonarea curbei 0 de la graficul IOPS al discului FLASH la valoarea de 30000 apare din cauza limitării benzii de comunicație a interfeței sau este o limitare dată intrinsec de valoarea latenței minime a discului.

În pagina următoare sunt prezentate rezultatele testului menționat anterior. Acest test trebuie în principiu rafinat prin configurare cu valori care să simuleze cât mai aproape comportarea mediului curent de lucru în diverse situații:

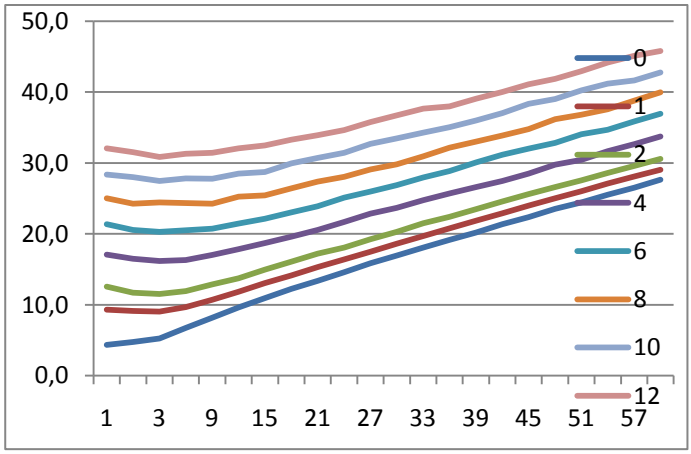
- Funcționare normală
- Regim de backup

Grafice teste

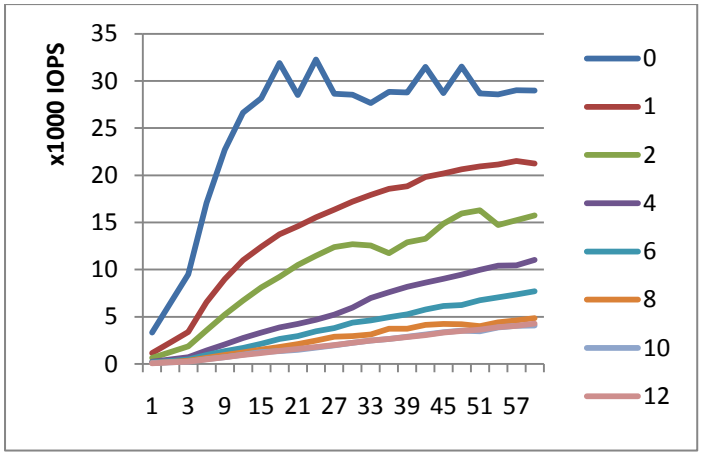
FLASH latență (ms)



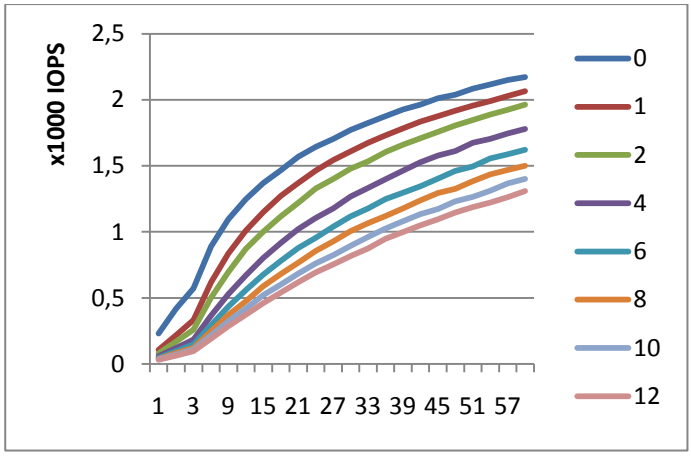
RAID 5 latență (ms)



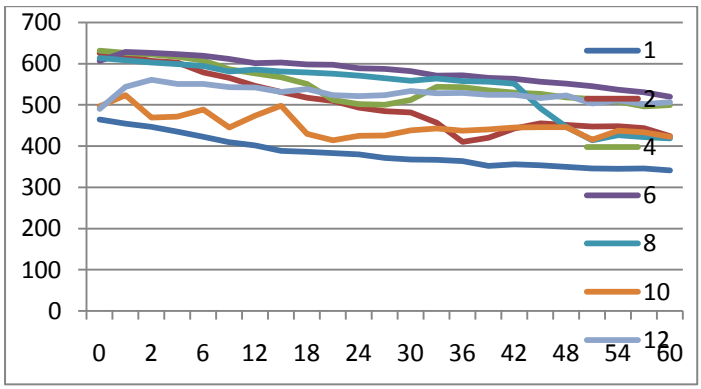
FLASH (mii IOPS)



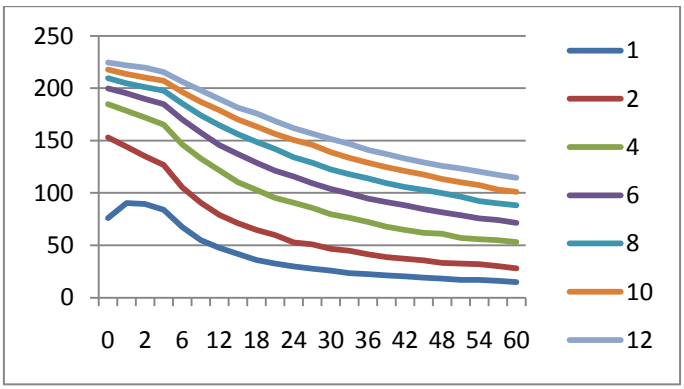
RAID 5 (mii IOPS)



FLASH (MBPS)



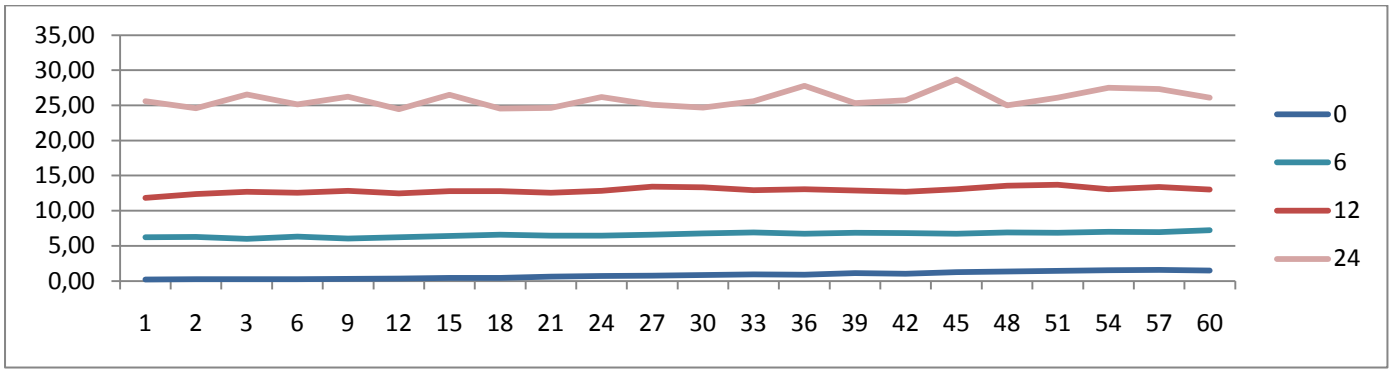
RAID5 (MBPS)



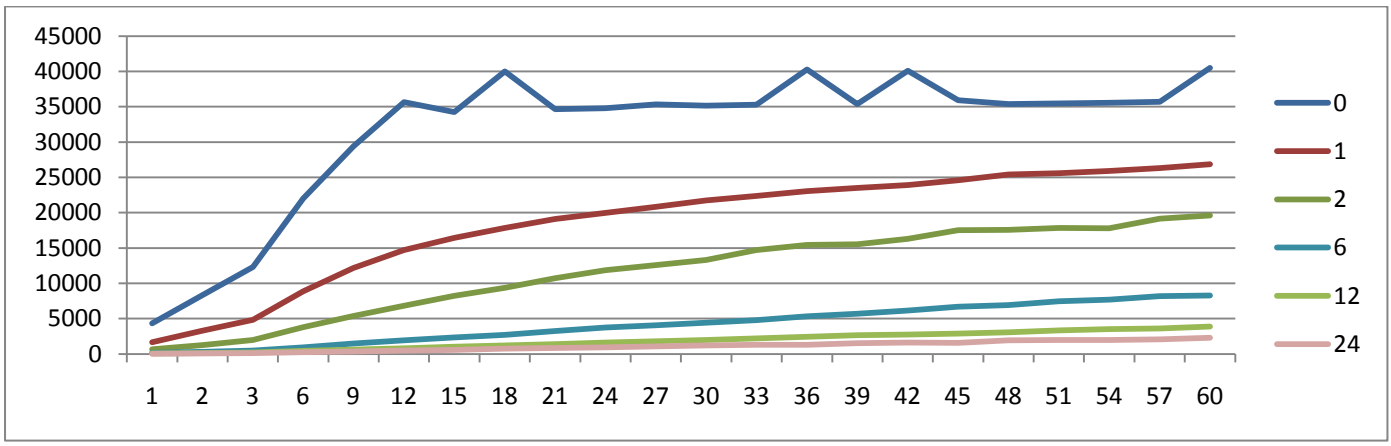
Mai jos sunt prezentate graficele Latență, IOPS şi MBPS pentru sistemul FLASH stresat în condiții mai grele de lucru:

- Tranzacțiile mici sunt de numai 4K (cât este block size definit la formatarea discului)
- S-a introdus un procent de 30% de tranzacții tip scriere pentru care fişa tehnică a sistemului de stocare oferă o cifră maximă IOPS mai mică

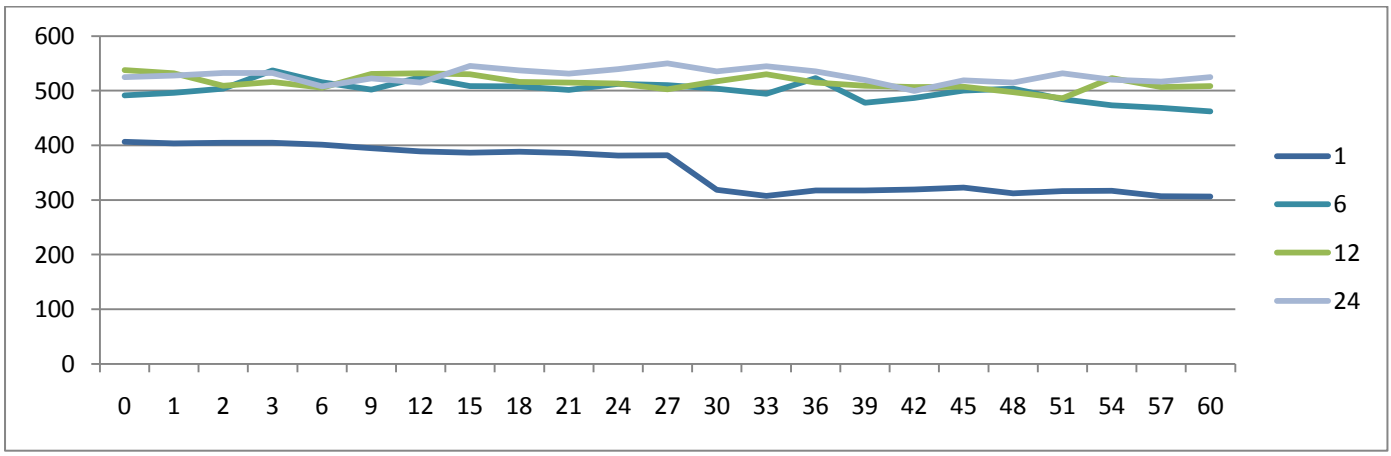
FLASH Latență (ms)



FLASH (IOPS)



FLASH (MBPS)



Contact

Richard Consulting SRL vă stă la dispoziție cu servicii de analiză, testare, evaluare şi modelare a sistemelor de stocare pentru aplicațiile dumneavoastra.

Vă rugam să ne contactați în acest sens la:

Richard Consulting SRL

Iancu Jianu 21, Otopeni
075100, Romania

http://www.richardconsulting.ro

tel. +40-21-3005550
fax +40-21-3005552
mobil: +40-722-317317

richard.vencu@richardconsulting.ro