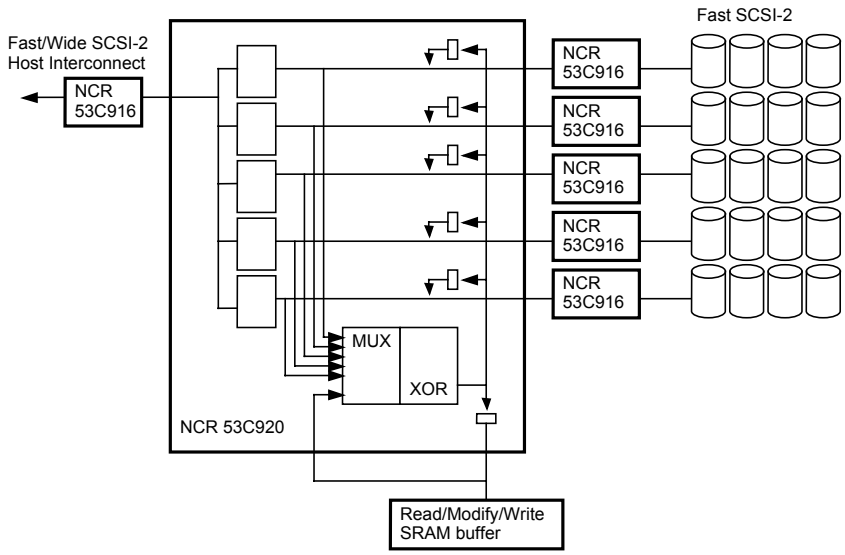
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΤΕΧΝΟΛΟΓΙΑ ΔΕΥΤΕΡΕΥΟΥΣΑΣ ΑΠΟΘΗΚΕΥΣΗΣ

RAID



ΓΚΕΖΕΡΛΗΣ ΒΕΛΙΣΣΑΡΗΣ, (Μ 4)
ΧΑΤΖΗΕΥΘΥΜΙΑΔΗΣ ΣΤΑΘΗΣ, (Μ 9)

ΑΘΗΝΑ,
ΙΟΥΛΙΟΣ 1995

---

*Εργασία στα πλαίσια του μαθήματος **Απόδοση Συστημάτων - Προσομοίωση** του Μεταπτυχιακού Προγράμματος Σπουδών.*

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ

Η παρουσίαση αυτή αντιμετωπίζει ζητήματα που σχετίζονται άμεσα με την τεχνολογία **RAID**, (**Redundant Arrays of Inexpensive Disks**). Το ενδιαφέρον στην συγκεκριμένη τεχνολογία έχει τελευταία αυξηθεί σημαντικά εξαιτίας των ραγδαίων εξελίξεων οι οποίες παρατηρούνται στο χώρο της τεχνολογίας μικροεπεξεργαστών και ολοκληρωμένων κυκλωμάτων. Γρηγορότεροι μικρο-επεξεργαστές και μεγαλύτερα συστήματα βασικής - κύριας μνήμης απαιτούν ισχυρότερα συστήματα περιφερειακής αποθήκευσης, με διαρκώς αυξανόμενες απαιτήσεις επιδόσεων. Οι εξελίξεις στο χώρο των μικροεπεξεργαστών εκτιμάται ότι θα προκαλέσουν αμελητέες βελτιώσεις στην ταχύτητα των υπολογιστικών συστημάτων αν δεν συνοδεύονται από ανάλογη βελτίωση στις επιδόσεις των συστημάτων περιφερειακής αποθήκευσης. Οι ρυθμοί με τους οποίους εξελίσσονται τα συστήματα περιφερειακής αποθήκευσης υπολείπονται σημαντικά των αντιστοίχων των μικροεπεξεργαστών, (τεχνολογία RISC κλπ.). Τέλος, η ραγδαία εξέλιξη των μικροεπεξεργαστών κατέστησε εφικτή την ανάπτυξη και εισαγωγή νέων τύπων εφαρμογών όπως οι εφαρμογές πολυμέσων, video, hypertext κλπ. οι οποίες απαιτούν ταχύτατη πρόσβαση σε εκτενή σύνολα δεδομένων, σε περιβάλλοντα πολλαπλών χρηστών και δικτύων υπολογιστών.

Οι συστοιχίες δίσκων, (disk arrays), και η τεχνολογία RAID αποτελούν μία διέξοδο στο πρόβλημα που επισημάνθηκε στην προηγούμενη παράγραφο. Τα διάφορα επίπεδα της τεχνικής RAID, για τα οποία εκτενής λόγος θα γίνει στις επόμενες παραγράφους, προτάθηκαν για πρώτη φορά το 1988 από τους ερευνητές Patterson, Gibson και Katz του Πανεπιστημίου της California, Berkeley. Η σχετική δημοσίευση, [PAT88], που είχε τον τίτλο "**A Case for Redundant Arrays of Inexpensive Disks (RAID)**", επιχειρούσε μία σύγκριση μεταξύ των τεχνικών RAID και SLED, (Single Large Expensive Disk), καθώς επίσης και την καθιέρωση των πέντε επιπέδων RAID, (1-5), κάνοντας εκτενή αναφορά στα συγκριτικά πλεονεκτήματα-μειονεκτήματα που εμφανίζουν.

Η βασική ιδέα πάνω στην οποία βασίζονται οι συστοιχίες δίσκων είναι η κατανομή των δεδομένων σε πολλαπλές συσκευές δίσκων. Μέσω της παράλληλης προσπέλασης τους επιτυγχάνονται μεγαλύτεροι ρυθμοί μεταφοράς δεδομένων, (data transfer rates), και ισχυρότερη ρυθμοαπόδοση, (throughput). Ομως, οι μεγάλες συστοιχίες δίσκων εμφανίζουν αδυναμίες όσον αφορά την διαθεσιμότητα τους. Μία συστοιχία με 100 δίσκους είναι 100 φορές πιθανότερο να υποστεί βλάβη από ένα μεμονωμένο δίσκο. Ο μέσος χρόνος μεταξύ βλαβών, (MTTF), ο οποίος για ένα μεμονωμένο δίσκο ανέρχεται στα 23 έτη, (200,000 ώρες), μειώνεται στις 2,000 ώρες

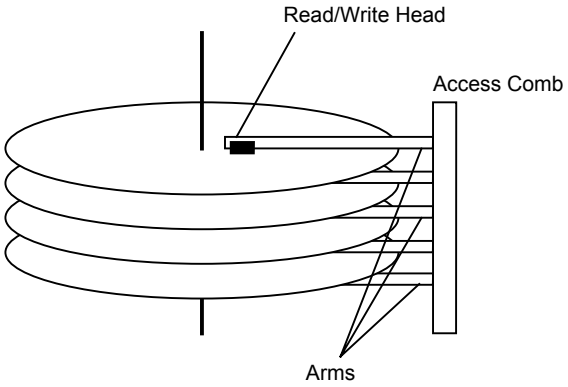
για μία συστοιχία με 100 δίσκους. Το πρόβλημα της αξιοπιστίας αντιμετωπίζεται με την εφαρμογή μεθόδων ανίχνευσης και αποκατάστασης λαθών, (Error detection & correction), οι οποίες όμως οδηγούν σε υποβάθμιση των επιδόσεων των σχετικών διατάξεων.

Οπως διαπιστώνεται παραπάνω, η εισαγωγή των συστοιχιών δίσκων και της τεχνολογίας RAID προσπαθεί αφενός μεν να καλύψει το χάσμα επιδόσεων μεταξύ των μηχανισμών δευτερεύουσας αποθήκευσης και των μικροεπεξεργαστών, κύριας μνήμης αφετέρου δε να κρατήσει σε υψηλά επίπεδα την διαθεσιμότητα και την ασφάλεια των δεδομένων.

Στην παράγραφο αυτή περιγράφεται συνοπτικά η δομή της παρουσίασης. Στο κεφάλαιο 2 πραγματοποιείται μία εκτενής αναφορά σε τεχνικά στοιχεία συστημάτων δίσκων και συστοιχιών δίσκων, (disk arrays). Στο κεφάλαιο 3 επιχειρείται μία εισαγωγή στα εφτά επίπεδα RAID τα οποία έχουν μέχρι στιγμής καθιερωθεί. Στο κεφάλαιο 4 παρουσιάζονται στοιχεία επιδόσεων για τα διάφορα επίπεδα RAID, στοιχεία κόστους κλπ. Στο κεφάλαιο 5 εξετάζονται ορισμένα ειδικά ζητήματα υλοποίησης συστοιχιών. Στο κεφάλαιο 6 έχουν συμπεριληφθεί νέες εξελίξεις αναφορικά με την τεχνολογία RAID καθώς και βελτιώσεις του επιπέδου RAID 5. Στο κεφάλαιο 7 πραγματοποιούνται αναφορές σε εφαρμογές της τεχνολογίας RAID με έμφαση σε εμπορικά διαθέσιμα προϊόντα. Τέλος, στο κεφάλαιο 8 πραγματοποιείται μία διερεύνηση του κόστους εξοπλισμού δευτερεύσουσας αποθήκευσης σε σχέση με την διαθεσιμότητα δεδομένων την οποία μπορεί αυτός ο εξοπλισμός να παρέχει.

2. ΤΕΧΝΙΚΑ ΣΤΟΙΧΕΙΑ ΔΙΣΚΩΝ & ΣΥΣΤΟΙΧΙΩΝ

Μία συσκευή δίσκου αποτελείται από ένα σύνολο μεταλλικών κυκλικών και σκληρών επιφανειών οι οποίες έχουν επενδυθεί με μαγνητικό υλικό και περιστρέφονται με σταθερή γωνιακή ταχύτητα γύρω από τον άξονα τους. Η διάμετρος τους ποικίλει από 1.3" μέχρι και 18". Κατά μήκος της ακτίνας των κυκλικών επιφανειών μετακινείται ένα σύνολο μαγνητικών κεφαλών ανάγνωσης-εγγραφής. Οι κεφαλές αυτές έχουν τοποθετηθεί στα άκρα βραχιόνων, (arms), οι οποίοι περιστρέφονται με την υποστήριξη ενός παράλληλου προς τον άξονα των κυκλικών επιφανειών στελέχους, (access comb). Σχηματικά η διάταξη του δίσκου εμφανίζεται στο ακόλουθο διάγραμμα, (Σχήμα 1).



Σχήμα 1: Οργάνωση Δίσκου

Παρά το γεγονός ότι όλες οι κεφαλές του συστήματος μετακινούνται ταυτόχρονα και βρίσκονται πάντα σε ίση απόσταση από το κέντρο του δίσκου, μόνο μία μπορεί να διαβάσει ή να γράψει σε κάποια συγκεκριμένη χρονική στιγμή. Τα δεδομένα που έχουν τοποθετηθεί σε μία περιφέρεια σταθερής ακτίνας αποτελούν μία άτρακτο, (track). Οι άτρακτοι όλων των κυκλικών επιφανειών οι οποίες βρίσκονται στην ίδια ακτίνα συγκροτούν ένα κύλινδρο, (cylinder). Οι άτρακτοι χωρίζονται σε τομείς, (sectors) ή σελίδες, (pages, clusters, blocks).

Τα μετρήσιμα μεγέθη που συνήθως καθορίζουν την επίδοση ενός συστήματος δίσκου, είναι ο χρόνος αναζήτησης, (seek time), ο χρόνος περιστροφής, (rotational latency time) και ο χρόνος μεταφοράς δεδομένων, (data transfer time). Ο χρόνος αναζήτησης αναφέρεται στην τοποθέτηση των κεφαλών πάνω από την σωστή άτρακτο για την ανάγνωση-εγγραφή δεδομένων. Ο χρόνος αυτός συνήθως κυμαίνεται από 1 έως 30 milliseconds, [CHE94], ανάλογα με την απόσταση που πρέπει να καλυφθεί και τον δίσκο που διαθέτουμε. Ως χρόνος περιστροφής θεωρείται ο χρόνος που απαιτείται για τον κατάλληλο τομέα, ύστερα από περιστροφή του όλου συστήματος, ώστε να τοποθετηθεί αυτός κάτω από την κεφαλή. Οι χρόνοι πλήρους περιστροφής ενός συστήματος δίσκου κυμαίνονται μεταξύ 8 και 28 milliseconds. Ο χρόνος μεταφοράς δεδομένων εξαρτάται άμεσα από τον ρυθμό με τον οποίο δεδομένα μεταφέρονται από και προς την μαγνητική επιφάνεια του δίσκου και είναι συνάρτηση της απόστασης της κεφαλής από το κέντρο της κυκλικής διάταξης, της πυκνότητας του μαγνητικού υλικού καθώς και του ρυθμού περιστροφής. Τυπικοί ρυθμοί μεταφοράς δεδομένων κυμαίνονται από 1 έως και 5 Mbytes/sec.

Οι εφαρμογές που χαρακτηρίζονται από μεγάλες απαιτήσεις σε λειτουργίες εισόδου-εξόδου, (I/O intensive applications), κατηγοριοποιούνται ως εξής:

- εφαρμογές υψηλού ρυθμού δεδομένων, (high data rate applications): ελάχιστες μετακινήσεις-τοποθετήσεις της κεφαλής. Εκτενείς και σειριακές προσπελάσεις.
- εφαρμογές υψηλού ρυθμού εισόδων/εξόδων, (high I/O rate applications): πολλαπλές μετακινήσεις της κεφαλής εξαιτίας σύντομων προσπελάσεων με μεγάλο βαθμό τυχαιότητας.

Οι επιστημονικές εφαρμογές που διαχειρίζονται μεγάλους πίνακες από δεδομένα τοποθετούνται στην πρώτη κατηγορία ενώ οι εφαρμογές επεξεργασίας δοσοληψιών, (transaction processing), στην δεύτερη.

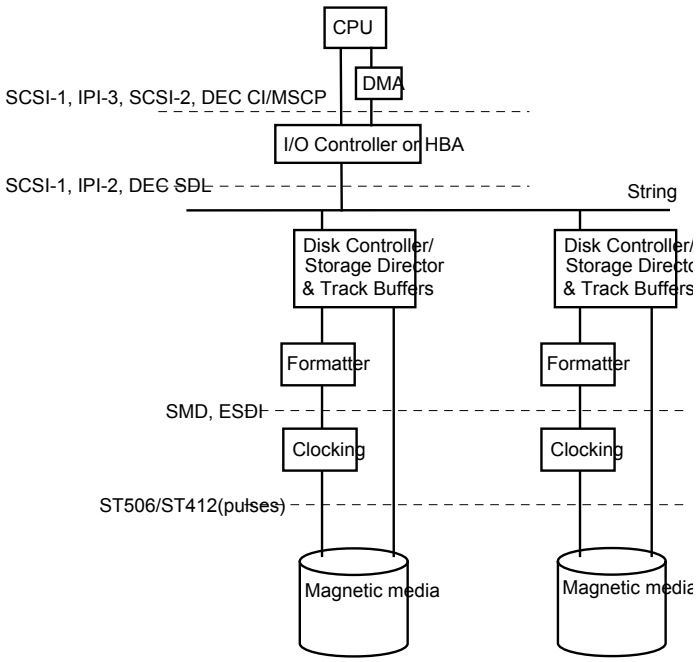
2.1 ΜΕΤΑΦΟΡΑ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΗΝ ΜΑΓΝΗΤΙΚΗ ΕΠΙΦΑΝΕΙΑ ΣΤΗΝ ΚΥΡΙΑ ΜΝΗΜΗ

Για την μεταφορά των δεδομένων από και προς την μαγνητική επιφάνεια του συστήματος δίσκου έχει οριστεί μία ιεραρχία από πρότυπα. Τα πρότυπα αυτά καλύπτουν την πλήρη επικοινωνία δίσκου-εφαρμογών. Στην επιφάνεια των δίσκων η πληροφορία αποτυπώνεται μέσω μεταβολών της μαγνητικής ροής, (flux reversals). Οι μεταβολές αυτές συλλαμβάνονται, ενισχύονται και μετατρέπονται σε ψηφιακό σήμα, (παλμοί). Διαδεδομένα πρωτόκολλα για το κατώτερο αυτό επίπεδο είναι τα ST506, ST412. Στο ακριβώς επόμενο επίπεδο, τα bits δεδομένων διαχωρίζονται από την πληροφορία συγχρονισμού. Στο επίπεδο αυτό ως πρότυπα αναφέρονται τα ESDI, (Enhanced Small Device Interface) και SMD, (Storage Module Interface).

Ακολούθως, τα bits δεδομένων ομαδοποιούνται σε bytes, εφαρμόζονται πάνω τους οι κατάλληλοι κώδικες για την διόρθωση λαθών και προωθούνται, ως blocks δεδομένων, στην βασική υπολογιστική διάταξη μέσω ενός προσαρμοστικού, (bus peripheral interface), όπως το SCSI, (Small Computer Standard Interface), ή το IPI-3, (το τρίτο επίπεδο του Intelligent Peripheral Interface). Τα πρότυπα αυτά καλύπτουν και την απεικόνιση των λογικών διευθύνσεων, (που χρησιμοποιούνται από τον υπολογιστή), σε φυσικές, (sector, cylinder, track). Η ενσωμάτωση του μηχανισμού απεικόνισης διευθύνσεων στο επίπεδο αυτό δίνει την δυνατότητα στο προσαρμοστικό να αποφεύγει, κατά τρόπο διαφανή, τις προβληματικές περιοχές του δίσκου και να επανατοποθετεί τα δεδομένα κατάλληλα.

Οι συσκευές που παρεμβάλλονται μεταξύ του δίσκου και της υπολογιστικής διάταξης για την μεταφορά των δεδομένων καθώς οι τοπολογίες με τις οποίες αυτές οργανώνονται ποικίλουν ανάλογα με το μέγεθος και τον τύπο του υποσυστήματος

εισόδου-εξόδου. Για την μεταφορά δεδομένων από ένα σύνολο δίσκων προς την κυρία μνήμη συχνά προβλέπεται η χρήση ενός κοινού διαύλου και ενός προσαρμοστή, (HBA, host bus adapter) όπως οι προσαρμοστές SCSI. Από τον προσαρμοστή τα δεδομένα μεταφέρονται στην κύρια μνήμη, (στους buffers του λειτουργικού συστήματος), με την τεχνική DMA μέσω του διαύλου συστήματος, (system bus), που ακολουθεί τυποποιήσεις όπως οι VME, (Versa Module Eurocard), S-Bus, EISA, (Extended Industry Standard Architecture) ή PCI, (Peripheral Component Interconnect).



Σχήμα 2: Σύνδεση δίσκου-κεντρικής μονάδας επεξεργασίας

2.2 ΕΞΕΛΙΞΗ ΤΗΣ ΤΕΧΝΟΛΟΓΙΑΣ ΔΙΣΚΩΝ

Ορισμένα από τα χαρακτηριστικά των σκληρών δίσκων εξελίσσονται με ταχείς ρυθμούς σε αντίθεση με άλλα που παραμένουν σχετικά στάσιμα. Αυτή την διαπίστωση δικαιολογεί ο πίνακας 1 που ακολουθεί:

Πίνακας 1: Εξέλιξη στην τεχνολογία δίσκων

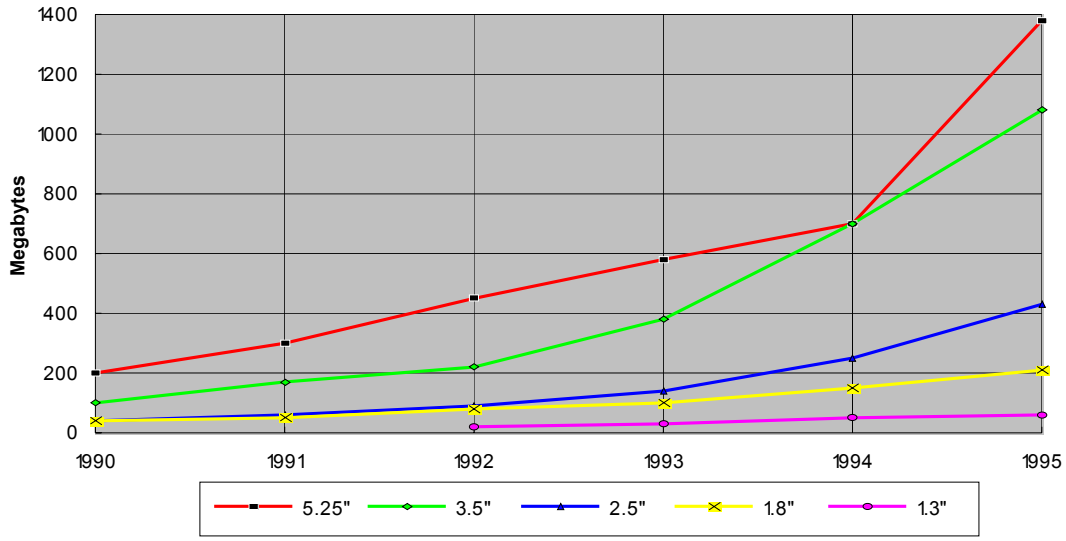
	1993	**Ρυθμός ανάπτυξης**
Χωρική πυκνότητα	50-150 Mbits ανά τετρ.ίντσα	27% ανά έτος
Γραμμική πυκνότητα	40,000-60,000 bits ανά ίντσα	13% ανά έτος
Πυκνότητα Ατράκτων	1,500-3,000 άτρακτοι ανά ίντσα	10% ανά έτος
Χωρητικότητα, (3.5")	100-2000 MB	27% ανά έτος

Ρυθμός Μεταφοράς	3-4 MB/s	22% ανά έτος
Χρόνος αναζήτησης	7-20 msec	8% ανά έτος

Οι μικρότερες αποστάσεις την μαγνητικής κεφαλής από την επιφάνεια του δίσκου καθώς και η εξέλιξη στην τεχνολογία του μαγνητικού υλικού έδωσαν μία ραγδαία ώθηση στην πυκνότητα εγγραφής των συσκευών. Η αύξηση της πυκνότητας εγγραφής επηρέασε τις συσκευές σε δύο βασικά σημεία.

Πρώτον, επέτρεψε στις χωρητικότητες δίσκων να παραμείνουν στάσιμες ή ακόμη και να αυξηθούν παρά το γεγονός ότι οι διαστάσεις τους μειώθηκαν από 5.25" το 1983 σε 1.3" το 1993. Η χρονική εξέλιξη της χωρητικότητας μίας μεμονωμένης μαγνητικής επιφάνειας, (single platter capacity), σε σχέση με τις διαστάσεις της, από το 1990 μέχρι το 1995 παρουσιάζεται στο διάγραμμα που ακολουθεί, (Σχήμα 3).

Σχήμα 3: Εξέλιξη της χωρητικότητας μίας μαγνητικής επιφάνειας



Source: Dataquest 1993

Δεύτερον, η αυξημένη πυκνότητα εγγραφής σε συνδυασμό με την αύξηση της ταχύτητας περιστροφής επέτρεψαν μία σημαντική άνοδο του ρυθμού μεταφοράς, (transfer rate). Σε αντίθεση με τα παραπάνω οι χρόνοι αναζήτησης βελτιώθηκαν ελάχιστα, και μειώθηκαν από 20 msec το 1980 σε 10 msec σήμερα. Οι χρόνοι περιστροφής ακολούθησαν ανάλογους αργούς ρυθμούς εξέλιξης, από 3600 rpm το 1980 σε 5400-7200 rpm σήμερα.

2.3 ΑΡΧΕΣ ΣΥΣΤΟΙΧΙΩΝ ΔΙΣΚΩΝ

Στα συστήματα συστοιχιών δίσκων υλοποιούνται δύο βασικές έννοιες, αυτές της κατανομής των δεδομένων σε πολλαπλές συσκευές, (data stripping), και της εισαγωγής πλεονάζουσας πληροφορίας, (redundancy). Η κατανομή των δεδομένων οδηγεί στην βελτίωση των επιδόσεων στο υποσύστημα εισόδου/εξόδου ενώ η εισαγωγή πλεονάζουσας πληροφορίας στην επαύξηση της διαθεσιμότητας του.

Η κατανομή των δεδομένων επιτρέπει, στην περίπτωση που υποβληθούν πολλαπλές και ανεξάρτητες μεταξύ τους αιτήσεις, να εξυπηρετηθούν παράλληλα από διαφορετικές συσκευές δίσκων. Η παράλληλη εκτέλεση ελαττώνει τον χρόνο αναμονής, (queuing time), των αιτήσεων για είσοδο/έξοδο. Απλές αιτήσεις που αφορούν πολλαπλά blocks, (multiple block), μπορούν να εξυπηρετηθούν από την συστοιχία με κατάλληλο συντονισμό των δίσκων που την συνθέτουν. Ετσι, επιτυγχάνονται υψηλότεροι ρυθμοί στην μεταφορά δεδομένων, (transfer rates). Συνοψίζοντας, η αύξηση του πλήθους των δίσκων στην συστοιχία επιφέρει σημαντική βελτίωση στις επιδόσεις που την διακρίνουν. Ταυτόχρονα όμως, υποβιβάζει αισθητά την αξιοπιστία του συστήματος. Σε συστοιχίες που αποτελούνται από 100 δίσκους η αξιοπιστία ανέρχεται στο 1/100 αυτής του ενός δίσκου.

Βασικά χαρακτηριστικά μίας συστοιχίας είναι (α) το μέγεθος των blocks με το οποίο πραγματοποιείται η κατανομή των δεδομένων στους δίσκους της, (η σχετική παράμετρος αναφέρεται ως **granularity of data interleaving**) και (β) η μέθοδος υπολογισμού της πλεονάζουσας πληροφορίας και τοποθέτησης της στην συστοιχία.

Ως προς το πρώτο χαρακτηριστικό οι συστοιχίες μπορούν να χαρακτηριστούν ως **fine** ή **coarse grained**. Οι fine grained συστοιχίες κατανέμουν τα δεδομένα στους επιμέρους δίσκους σε σχετικά μικρά blocks-μονάδες πληροφορίας. Οι αιτήσεις I/O, ασχέτως του μεγέθους που αφορούν, πρέπει να προσπελάσουν ένα μεγάλο πλήθος δίσκων, (πιθανά όλους). Το θετικό αποτέλεσμα της στρατηγικής αυτής είναι οι αρκετά υψηλοί ρυθμοί μεταφοράς δεδομένων, (transfer rates), ενώ τα αρνητικά αποτελέσματα αναφέρονται πρώτον στην αδυναμία διεκπεραίωσης περισσότερων της μίας αιτήσεων ταυτόχρονα και δεύτερον στην δαπάνη χρόνου σε κάθε δίσκο για την τοποθέτηση των κεφαλών σε κάθε αίτηση.

Οι coarse grained συστοιχίες κατανέμουν την πληροφορία σε μεγάλα blocks. Περιορισμένες αιτήσεις I/O επιβάλλουν την προσπέλαση σε ένα μικρό πλήθος δίσκων. Κατά συνέπεια, πολλαπλές περιορισμένες σε μέγεθος αιτήσεις μπορούν να ικανοποιηθούν ταυτόχρονα ενώ και στην περίπτωση των μεγαλύτερων αιτήσεων

υπάρχει το πλεονέκτημα του υψηλού ρυθμού μεταφοράς, όπως αυτό επισημάνθηκε στην περίπτωση των fine grained.

Για τον υπολογισμό της πλεονάζουσας πληροφορίας ασφαλείας, στην συστοιχία, χρησιμοποιούνται, κατά κύριο λόγο, μέθοδοι-τεχνικές όπως η ισοτιμία, (parity), καθώς και οι κώδικες Hamming ή Reed-Solomon. Για την τοποθέτηση της πληροφορίας αυτής στους δίσκους της συστοιχίας ακολουθούνται δυο στρατηγικές. Η πρώτη από τις στρατηγικές αυτές προβλέπει την τοποθέτηση της πλεονάζουσας πληροφορίας σε ένα περιορισμένο πλήθος δίσκων ενώ η δεύτερη την ομοιόμορφη κατανομή της σε όλους τους δίσκους που συνθέτουν την συστοιχία. Η τελευταία κρίνεται πλεονεκτικότερη γιατί επιτρέπει σημαντική εξισορρόπηση του φόρτου του συστήματος, (load balancing).

3. ΤΑ ΕΠΙΠΕΔΑ RAID

Στις παραγράφους που ακολουθούν αναλύονται τα 5 επίπεδα RAID τα οποία προτάθηκαν στην δημοσίευση [PAT88], (1 έως 5), καθώς και 2 ακόμη επίπεδα, (τα 0 και 6), τα οποία έχουν καθιερωθεί κατά την εξέλιξη της σχετικής τεχνολογίας.

3.1 RAID ΕΠΙΠΕΔΟ 0

Στο RAID επίπεδο 0 δεν προβλέπεται ο υπολογισμός και η διατήρηση πλεονάζουσας πληροφορίας, (**nonredundant disk array**). Τα συστήματα επιπέδου 0 εμφανίζουν τις καλύτερες επιδόσεις εγγραφής, (write performance), επειδή, όπως αναφέρθηκε, δεν απαιτούν τη ενημέρωση πληροφορίας ασφαλείας. Τα δεδομένα κατανέμονται σε πολλαπλούς δίσκους, (disk stripping). Κάθε φορά που κάποιες πληροφορίες πρέπει να γραφτούν στην συστοιχία, το πρώτο segment τους αποθηκεύεται στον δίσκο x, το δεύτερο στον x+1 κλπ. Όταν απαιτείται μεταφορά δεδομένων, τα τμήματα των πληροφοριών βρίσκονται σε διαφορετικούς δίσκους και κατά συνέπεια η άντληση τους μπορεί να πραγματοποιηθεί σε επικαλυπτόμενα χρονικά διαστήματα, (overlapped I/O).

Το επίπεδο 0 δεν εμφανίζει υψηλές επιδόσεις κατά την ανάγνωση από την συστοιχία, σε αντίθεση με σχήματα όπως η διατήρηση αντιγράφου, (data duplication, mirroring). Σε συστήματα mirroring είναι εφικτή η δρομολόγηση των αιτήσεων στους συγκεκριμένους δίσκους που πρόκειται να εμφανίσουν μικρότερη καθυστέρηση αναζήτησης και περιστροφής. Η απουσία πλεονάζουσας πληροφορίας οδηγεί, μετά από δυσλειτουργία ενός δίσκου σε απώλεια δεδομένων, (data loss).

Οι συστοιχίες δίσκων που δεν διαθέτουν πλεονάζουσα πληροφορία, (nonredundant arrays), χρησιμοποιούνται σε περιβάλλοντα supercomputers, γιατί στον χώρο αυτό οι επιδόσεις είναι σημαντικότερες της αξιοπιστίας.

3.2 RAID ΕΠΙΠΕΔΟ 1

Στο επίπεδο 1 υιοθετείται η διατήρηση πλήρους αντιγράφου των δίσκων του συστήματος, (**Disk mirroring** ή **Shadowing**), [BIT88]. Από τους διαθέσιμους δίσκους ένας αριθμός χρησιμοποιείται σαν ο κύριος τόπος αποθήκευσης των πληροφοριών ενώ σε ισάριθμους δίσκους διατηρείται, με ελάχιστη χρονική καθυστέρηση, ένα ακριβές, πλεονάζον αντίγραφο των ίδιων πληροφοριών. Οταν δεδομένα ανακτούνται από την συστοιχία, επιλέγονται οι δίσκοι με τους μικρότερους χρόνους αναμονής, περιστροφής και μεταφοράς, (ευνοϊκότερη θέση κεφαλών κλπ.). Αυτή είναι και η σημαντικότερη βελτίωση στην ταχύτητα της I/O διαδικασίας η οποία μπορεί να επιτευχθεί.

Σε περίπτωση που κάποιος από τους δίσκους της συστοιχίας υποστεί βλάβη, τα σχετικά δεδομένα λαμβάνονται **άμεσα** από τον εφεδρικό του και ο χρόνος αδράνειας του συστήματος, (down-time) είναι μηδενικός. Η τεχνική του mirroring χρησιμοποιείται για την υποστήριξη της λειτουργίας βάσεων δεδομένων γιατί η διαθεσιμότητα και ο υψηλός ρυθμός εκτέλεσης δοσοληψιών θεωρούνται ιδιαίτερα κρίσιμοι παράγοντες.

Μία συγγενής με το disk mirroring τεχνική είναι αυτή του duplexing, [ALF92]. Στο **disk duplexing** οι βασικοί δίσκοι προσαρτώνται σε διαφορετικούς ελεγκτές από ότι οι εφεδρικοί, (mirrored). Οι λειτουργίες ανάγνωσης και εγγραφής είναι πλέον εφικτό να εκτελούνται τελείως παράλληλα. Επίσης αποφεύγεται ο σχηματισμός ενός μεμονωμένου σημείου πιθανής βλάβης, (single point of failure), το οποίο, στην περίπτωση του disk mirroring, εντοπίζεται στον ελεγκτή των μονάδων δίσκου.

3.3 RAID ΕΠΙΠΕΔΟ 2

Στα συστήματα μνήμης είναι εφικτή η πλήρης αποκατάσταση της πληροφορίας, σε περίπτωση βλάβης, με σημαντικά μικρότερο κόστος από αυτό της διατήρησης πλήρους αντιγράφου. Στα συστήματα αυτά χρησιμοποιούνται **κώδικες**

Hamming, (ισοτιμία για επικαλυπτόμενα υποσύνολα δεδομένων κλπ.). Σε μία από τις δυνατές παραλλαγές του επιπέδου 2, απαιτούνται 3 δίσκοι για την διατήρηση της πλεονάζουσας πληροφορίας τεσσάρων. Ο αριθμός των πλεοναζόντων δίσκων στην συστοιχία είναι λογαριθμικά εξαρτώμενος από τον συνολικό αριθμό δίσκων. Κατά συνέπεια η αποδοτικότητα, (efficiency), όσον αφορά στην αποθήκευση αυξάνεται με το πλήθος των δίσκων.

Εάν ένας δίσκος υποστεί βλάβη, ορισμένα από τα στοιχεία ισοτιμίας θα έχουν ασυνεπείς τιμές, (inconsistent). Το συστατικό που έχει υποστεί βλάβη προσδιορίζεται μέσω της τομής των ασυνεπών συνόλων. Η χαμένη πληροφορία μπορεί να ανακτηθεί συνθέτοντας ένα σύνολο από τα υπόλοιπα στοιχεία, συμπεριλαμβανόμενου και του στοιχείου ισοτιμίας. Το bit το οποίο λείπει τίθεται σε κατάσταση 0 ή 1 έτσι ώστε να αποδίδει την σωστή ισοτιμία για το σύνολο.

Με τον τρόπο που περιγράφηκε παραπάνω πολλαπλοί πλεονάζοντες δίσκοι απαιτούνται για τον καθορισμό του προβληματικού δίσκου, αλλά μόνο ένας απαιτείται για την ανάκτηση της χαμένης πληροφορίας. Γενικά, η υλοποίηση συστημάτων επιπέδου 2, λόγω της πολυπλοκότητας που το χαρακτηρίζει, (κώδικες Hamming κλπ.), θεωρείται δύσκολη. Επίσης, το επίπεδο 2 δεν έχει πρακτική σημασία για συστήματα με πλήθος δίσκων μικρότερο από 10 γιατί το ποσοστό της χωρητικότητας που είναι διαθέσιμο για πραγματικά δεδομένα ελαττώνεται πολύ γρήγορα και μπορεί να πέσει κάτω από το 50%, [ALF92].

3.4 RAID ΕΠΙΠΕΔΟ 3

Μία σημαντική βελτίωση στον μηχανισμό που προβλέπεται από το επίπεδο 2 μπορεί να βασιστεί στην δυνατότητα των ελεγκτών σκληρών δίσκων να προσδιορίζουν, με σχετική ευκολία, τα ελαττωματικά συστατικά της συστοιχίας. Κατά συνέπεια, είναι εφικτή η χρήση μόνο **ενός δίσκου ισοτιμίας**, (parity disk), σε αντίθεση με το πλήθος των δίσκων που προβλέπονται από το επίπεδο 2.

Στο επίπεδο 3 εφαρμόζεται η τεχνική **bit-interleaved parity**. Τα δεδομένα κατανέμονται, ανά bit, στους επιμέρους δίσκους, και το bit ισοτιμίας αποθηκεύεται στον δίσκο ισοτιμίας, (parity disk). Η ύπαρξη του δίσκου ισοτιμίας επιτρέπει την ανοχή του συστήματος σε οποιαδήποτε μεμονωμένη βλάβη. Σύμφωνα με τα παραπάνω, οι αιτήσεις ανάγνωσης επιβάλλουν την προσπέλαση όλων των δίσκων δεδομένων, (data disks), ενώ οι αιτήσεις εγγραφής την προσπέλαση των δίσκων

δεδομένων καθώς και του δίσκου ισοτιμίας. Σε κάθε χρονική στιγμή μπορεί να εξυπηρετηθεί μόνο μία αίτηση I/O.

Εφόσον, ο δίσκος ισοτιμίας δεν μπορεί να συμμετάσχει στην διαδικασία ανάγνωσης, οι επιδόσεις του επιπέδου 3 στον τομέα της εγγραφής υπολείπονται των επιδόσεων άλλων συστοιχιών, όπου τα δεδομένα καθώς και η πληροφορία ισοτιμίας διασπείρονται σε όλους τους διαθέσιμους δίσκους. Το επίπεδο 3 είναι απλούστερο στην υλοποίηση από τα επίπεδα 4, 5 και 6. Χρησιμοποιείται για την υποστήριξη εφαρμογών που δεν απαιτούν μεγάλους ρυθμούς I/O.

3.5 RAID ΕΠΙΠΕΔΟ 4

Ο μηχανισμός του επιπέδου 4 είναι τελείως ανάλογος με αυτόν του επιπέδου 3. Η διαφορά τους εντοπίζεται στο γεγονός ότι η πληροφορία δεν κατανέμεται στους δίσκους ανά bit αλλά ανά block, (**block interleaved parity**), μη συγκεκριμένου μεγέθους. Το μέγεθος των blocks καλείται **stripping unit**. Οι αιτήσεις ανάγνωσης που αφορούν όγκο μικρότερο από το stripping unit προσπελαύνουν μόνο ένα δίσκο δεδομένων. Οι αιτήσεις εγγραφής θα πρέπει να ενημερώσουν κατάλληλα τους δίσκους δεδομένων αλλά και να υπολογίσουν και καταχωρήσουν την πληροφορία ισοτιμίας.

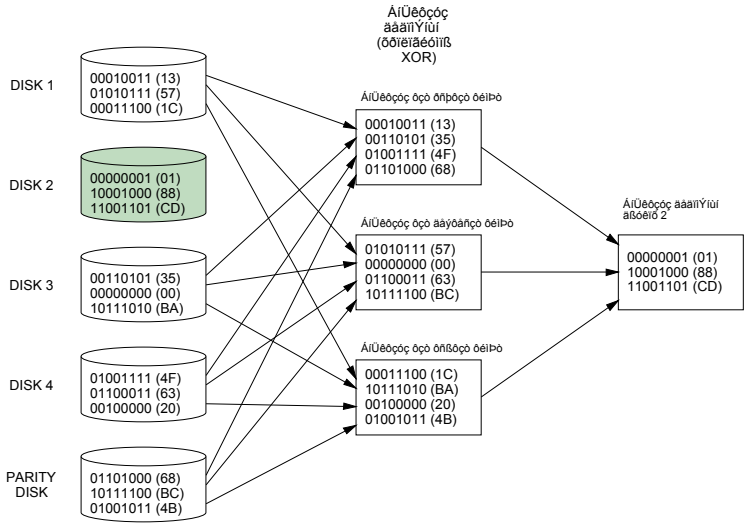
Μεγάλες εγγραφές που εκτείνονται σε όλους τους δίσκους δεδομένων επιτρέπουν τον άμεσο υπολογισμό της πληροφορίας ισοτιμίας εφαρμόζοντας απλά την λογική συνάρτηση XOR στα νέα δεδομένα. Για μικρότερες σε όγκο εγγραφές, που αφορούν μόνο ένα δίσκο, η πληροφορία ισοτιμίας μπορεί να ενημερωθεί εξετάζοντας τις διαφορές των νέων δεδομένων από τα παλιά και ενεργώντας κατάλληλα στο περιεχόμενο του δίσκου ισοτιμίας. Ετσι, για την πραγματοποίηση των εγγραφών μικρού μεγέθους ακολουθείται η διαδικασία read-modify-write, με 4 λειτουργίες I/O.

Επειδή, ο δίσκος ισοτιμίας πρέπει να ενημερωθεί σε κάθε εγγραφή, διαμορφώνεται μία στενωπός κίνησης, (bottleneck), η οποία υποβαθμίζει και τις γενικότερες επιδόσεις της συστοιχίας. Εξαιτίας αυτού ακριβώς του προβλήματος, που εμφανίζει το επίπεδο 4, προτιμάται η κατανομή των δεδομένων ισοτιμίας σε όλους τους δίσκους της συστοιχίας, όπως αυτή προβλέπεται από το επίπεδο 5.

Στο σχήμα 4 που ακολουθεί παρουσιάζεται η ανάκτηση δεδομένων, (data recovery), στα επίπεδα 3 ή 4. Η πληροφορία ισοτιμίας που αποθηκεύεται

προσδιορίζεται μέσω του λογικού XOR των δεδομένων. Ο σκιασμένος δίσκος έχει αντιμετωπίσει πρόβλημα, (δυστοκία). Τα δεδομένα του, τα οποία χάθηκαν, προσδιορίζονται από το λογικό XOR των υπολοίπων δίσκων καθώς και του parity.

Σχήμα 4: Ανάκτηση δεδομένων στα επίπεδα 3 ή 4



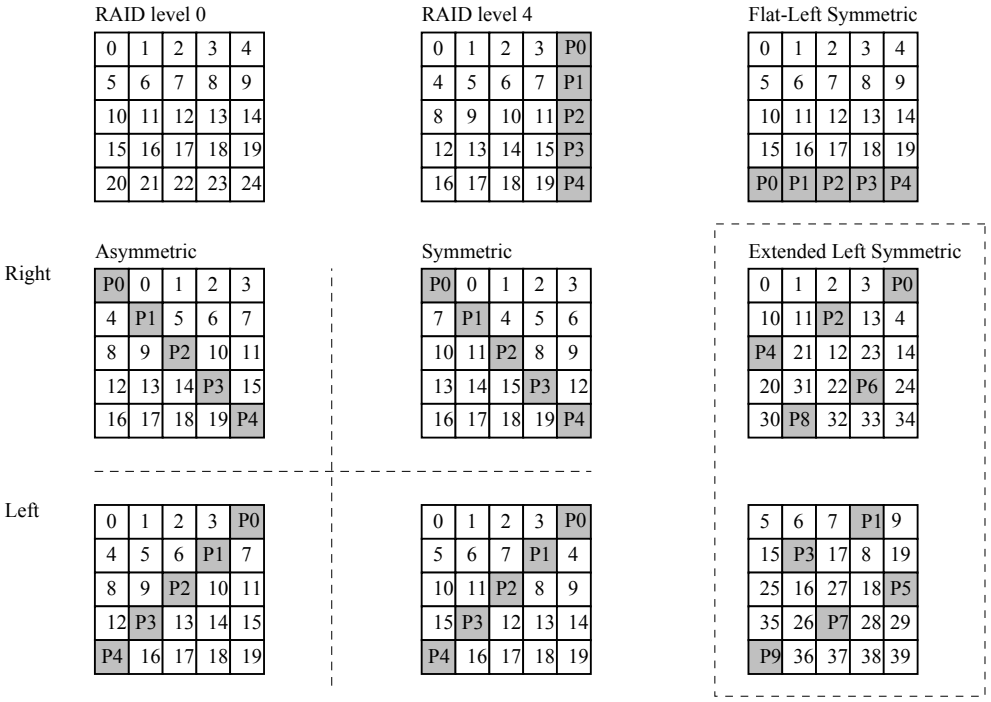
3.6 RAID ΕΠΙΠΕΔΟ 5

Η κατανομή των δεδομένων ισοτιμίας σε όλους τους δίσκους της συστοιχίας, (**block interleaved distributed parity**), επιφέρει πολλά θετικά αποτελέσματα για την λειτουργία της. Καταρχάς, αποφεύγεται η διαμόρφωση της στενωπού στο δίσκο ισοτιμίας, όπως αυτή παρατηρήθηκε στο επίπεδο 4. Επίσης τα δεδομένα κατανέμονται σε όλους τους δίσκους της συστοιχίας, σε αντίθεση με τα επίπεδα 3 και 4 που απασχολούσαν, για τον ίδιο λόγο, όλους τους δίσκους της συστοιχίας πλην ενός.

Το επίπεδο 5 εμφανίζει τις καλύτερες επιδόσεις, (σε σχέση με όλα τα υπόλοιπα επίπεδα που διατηρούν πλεονάζουσα πληροφορία), όσον αφορά στην εκτέλεση εγγραφών μεγάλου όγκου, (large writes), αναγνώσεων μικρού και μεγάλου όγκου, (small & large reads). Για τις εγγραφές μικρού όγκου απαιτείται λειτουργία read-modify-write για την ενημέρωση των δεδομένων ισοτιμίας.

Για την κατανομή των δεδομένων ισοτιμίας στους δίσκους της συστοιχίας ακολουθούνται διάφορες στρατηγικές, [LEE91], οι βασικότερες από τις οποίες παρουσιάζονται στο ακόλουθο διάγραμμα, (Σχήμα 5), συγκρινόμενες με τα επίπεδα 0 και 4.

Σχήμα 5: Τοποθέτηση πληροφορίας ισοτιμίας



Το κάθε μικρό τετράγωνο στο παραπάνω σχήμα αντιπροσωπεύει ένα striping unit ενώ η κάθε στήλη τετραγώνων την ίδια συσκευή δίσκου. Τα σκιασμένα τετράγωνα αντιπροσωπεύουν πληροφορία ισοτιμίας, (το P0 καλύπτει τα units 0, 1, 2 και 3, το P1 τα units 4, 5, 6 και 7 κλπ.).

Το σχήμα **right asymmetric** προέρχεται από το επίπεδο 0. Σε κάθε οριζόντιο επίπεδο αφαιρούνται stiping units δεδομένων επειδή εισάγεται πληροφορία ισοτιμίας. Διαδοχικά striping units ισοτιμίας εισέρχονται σε θέσεις, (στην μήτρα του σχήματος), που σταδιακά μετακινούνται προς τα δεξιά, (μία θέση σε κάθε γραμμή). Το σχήμα **left asymmetric** είναι το ίδιο με το right asymmetric με την διαφορά ότι οι θέσεις των stiping units ισοτιμίας σταδιακά μετακινούνται προς τα αριστερά.

Το σχήμα **right symmetric** προέρχεται από οριζόντια περιστροφή προς τα δεξιά, (κατά μία θέση σε κάθε γραμμή), ολοκλήρου του **stripe** ισοτιμίας, (τα data units μαζί με το parity unit που τα πλαισιώνει), όπως αυτό διαμορφώνεται στο επίπεδο 4. Το σχήμα **left symmetric** είναι το ίδιο με το right symmetric με την διαφορά ότι η μετακίνηση γίνεται προς τα δεξιά.

Από τις στρατηγικές τοποθέτησης των units ισοτιμίας οι οποίες περιγράφηκαν παραπάνω, τις καλύτερες επιδόσεις έχει επιδείξει η left symmetric.

Το σχήμα **extended left symmetric** προκύπτει από το επίπεδο 0, με την κάθετη προώθηση των data units καθώς εισάγονται parity units. Για διαδοχικά parity stripes, το σημείο εισαγωγής του parity unit μετακινείται κατά μία θέση προς τα αριστερά. Το σχήμα **flat left symmetric** προκύπτει από το extended left symmetric με την ομαδοποίηση της πληροφορίας ισοτιμίας και την τοποθέτηση της σε συγκεκριμένη θέση σε κάθε μονάδα δίσκου.

3.7 RAID ΕΠΙΠΕΔΟ 6

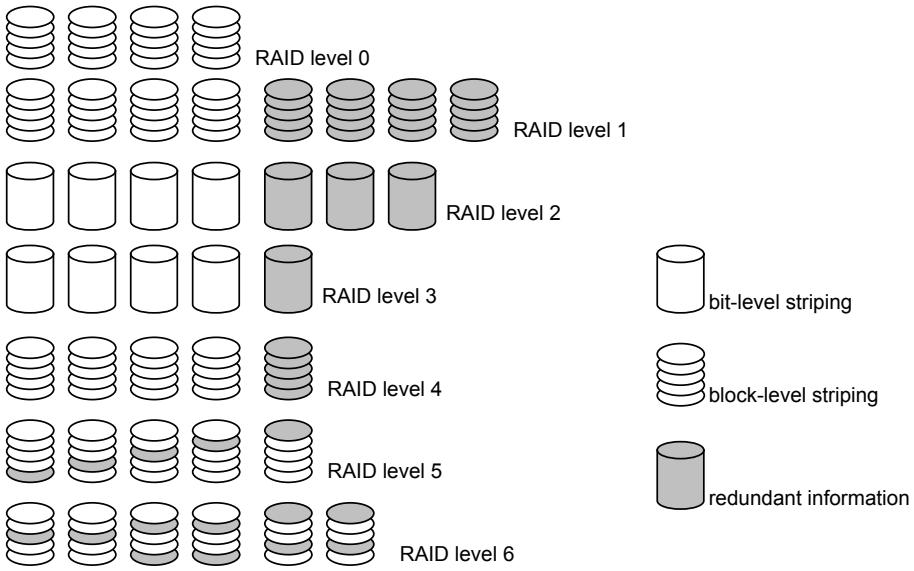
Το βασικότερο πρόβλημα αναφορικά με την πληροφορία ισοτιμίας, η οποία χρησιμοποιείται στα επίπεδα 1-5 είναι η δυνατότητα της συστοιχίας να αντιμετωπίσει μεμονωμένες βλάβες. Οσο μεγαλώνουν οι διαστάσεις των συστοιχιών αυξάνεται η πιθανότητα εμφάνισης πολλαπλών βλαβών και, κατά συνέπεια, απαιτούνται ισχυρότεροι κώδικες για την αντιμετώπιση του προβλήματος.

Ενα πλεονεκτικότερο σχήμα, όσον αφορά στην ασφάλεια των δεδομένων, είναι αυτό που προβλέπεται στο επίπεδο 6. Στο επίπεδο 6 γίνεται χρήση του κώδικα **Reed-Solomon** για την προστασία από διπλές βλάβες, με την χρήση, κατ' ελάχιστο, δύο δίσκων για την αποθήκευση πλεονάζουσας πληροφορίας, (redundant disks). Το σχήμα του επιπέδου 6 καλείται **P+Q**. Οι συστοιχίες που υιοθετούν το P+Q μοιάζουν σημαντικά με τις συστοιχίες block interleaved distributed parity, (επίπεδο 5), και λειτουργούν με τον ίδιο τρόπο.

Οι συστοιχίες P+Q χρησιμοποιούν την διαδικασία read-modify-write για την εκτέλεση εγγραφών περιορισμένου όγκου. Απαιτούν, για την παραπάνω διαδικασία 6 I/O προσπελάσεις, αντί για 4 που απαιτούνται στο επίπεδο 4, επειδή πρέπει να ενημερωθούν και οι δύο δίσκοι πλεονάζουσας πληροφορίας, (P και Q).

Το Σχήμα 6 που ακολουθεί παρουσιάζει μία συνολική εικόνα των επιπέδων RAID.

Σχήμα 6: Επίπεδα RAID 0 - 6



4. ΕΠΙΔΟΣΕΙΣ, ΣΥΓΚΡΙΣΕΙΣ ΚΟΣΤΟΥΣ ΚΑΙ ΑΞΙΟΠΙΣΤΙΑ

Οι τρεις κύριες μετρικές που χρησιμοποιούνται στην εκτίμηση της απόδοσης των συστοιχιών δίσκων είναι οι: αξιοπιστία, (reliability), επίδοση, (performance), και κόστος, (cost). Η χρήση και των τριών είναι σημαντική για την ορθή αξιολόγηση μίας συστοιχίας. Στις παραγράφους που ακολουθούν συγκρίνονται οι συστοιχίες επιπέδων RAID 0-6 με βάση τις προαναφερθήσες μετρικές.

4.1 ΚΑΝΟΝΕΣ ΓΙΑ ΤΟΝ ΚΑΘΟΡΙΣΜΟ ΤΩΝ ΜΕΤΡΙΚΩΝ ΕΠΙΔΟΣΗΣ - ΚΟΣΤΟΥΣ

Υπάρχουν πολλοί διαφορετικοί τρόποι υπολογισμού των παραπάνω μετρικών και ακόμη περισσότεροι για το πως μπορούν αυτοί να χρησιμοποιηθούν. Για παράδειγμα, θα πρέπει η επίδοση να υπολογίζεται σε αριθμό I/O ανά δευτερόλεπτο, σε bytes ανά δευτερόλεπτο, ή σε χρόνο απόκρισης; Η μήπως θα ήταν καταλληλότερη μία υβριδική μετρική όπως αριθμός I/O ανά δευτερόλεπτο ανά νομισματική μονάδα.

Η μέθοδος που θα χρησιμοποιηθεί θα πρέπει να βασίζεται κυρίως στον σκοπό της σύγκρισης καθώς και στην προβλεπόμενη χρήση του συστήματος. Ετσι, για τις εφαρμογές διαμοιραζόμενου χρόνου η καταλληλότερη μετρική είναι η συνολική χωρητικότητα που είναι διαθέσιμη στον χρήστη, (user capacity), ανά νομισματική μονάδα. Σε εφαρμογές επεξεργασίας δοσοληψιών είναι ο αριθμός των I/O ανά

δευτερόλεπτο ανά νομισματική μονάδα, ενώ σε επιστημονικές εφαρμογές είναι δυνατόν να είναι ο αριθμός των byte ανά δευτερόλεπτο ανά νομισματική μονάδα.

Οσον αφορά τα συστήματα δευτερεύουσας αποθήκευσης και ειδικότερα τις συστοιχίες δίσκων υπάρχει ένας σαφής προσανατολισμός προς το μέγεθος της ρυθμοαπόδοσης, (throughput). Δηλαδή η συνολική ρυθμοαπόδοση, (aggregate throughput), του συστήματος θεωρείται περισσότερο κρίσιμη από τον χρόνο απόκρισης, (response time), μίας συγκεκριμένης αίτησης. Επίσης, στα συστήματα αυτά οι επιδόσεις αυξάνονται γραμμικά με την προσθήκη νέων συστατικών. Για παράδειγμα, εάν ένας δίσκος μπορεί να εκτελέσει 30 λειτουργίες I/O το δευτερόλεπτο, η προσθήκη ενός νέου θα διπλασιάσει την απόδοση του συστήματος σε 60 λειτουργίες I/O το δευτερόλεπτο.

Για να συγκριθούν οι επιδόσεις των συστοιχιών δίσκων θα πρέπει να κανονικοποιηθεί η επίδοσή τους ως προς το κόστος τους. Ετσι, θα χρησιμοποιηθεί η μετρική που προσδιορίζεται από τον αριθμό των λειτουργιών I/O ανά δευτερόλεπτο, (throughput), ανά νομισματική μονάδα και όχι ο απόλυτος αριθμός λειτουργιών I/O το δευτερόλεπτο.

Για τις ανάγκες σύγκρισης συστοιχιών δίσκων επιλέγονται συστήματα ισοδύναμης χωρητικότητας αρχείων, (**file capacity**), όπου χωρητικότητα αρχείων θεωρείται το ποσό πληροφορίας που μπορεί να καταχωρηθεί από το σύστημα αρχείων, (file system), στην συστοιχία μη συμπεριλαμβανόμενης της πλεονάζουσας πληροφορίας. Η σύγκριση συστημάτων ίσης χωρητικότητας αρχείων διευκολύνει την επιλογή ισοδύναμων φόρτων για διαφορετικά σχήματα πλεονασμού.

Τέλος, πρέπει να αναφερθεί ότι υπάρχει, προς το παρόν, πολύ σύγχυση σχετικά με τις συγκρίσεις των επιπέδων RAID 1-5. Η σύγχυση αυτή οφείλεται στο γεγονός ότι το κάθε επίπεδο RAID δεν προσδιορίζει, σε ορισμένες περιπτώσεις, τη συγκεκριμένη υλοποίηση ενός συστήματος αλλά τον σχεδιασμό και τον τρόπο χρήσης του.

4.2 ΣΥΓΚΡΙΣΕΙΣ

Στον πίνακα 2 που ακολουθεί παρουσιάζεται η μέγιστη ρυθμοαπόδοση ανά νομισματική μονάδα, κανονικοποιημένη ως προς τις τιμές του επιπέδου RAID 0, για τα επίπεδα 0, 1, 3, 5 και 6.

Πίνακας 2: Ρυθμοαπόδοση ανά νομισματική μονάδα κανονικοποιημένη ως προς τις τιμές του RAID 0

Levels	Small Read	Small Write	Large Read	Large Write	Storage Efficiency
RAID 0	1	1	1	1	1
RAID 1	1	1/2	1	1/2	1/2
RAID 3	1/G	1/G	(G-1)/G	(G-1)/G	(G-1)/G
RAID 5	1	max(1/G, 1/4)	1	(G-1)/G	(G-1)/G
RAID 6	1	max(1/G, 1/6)	1	(G-2)/G	(G-2)/G

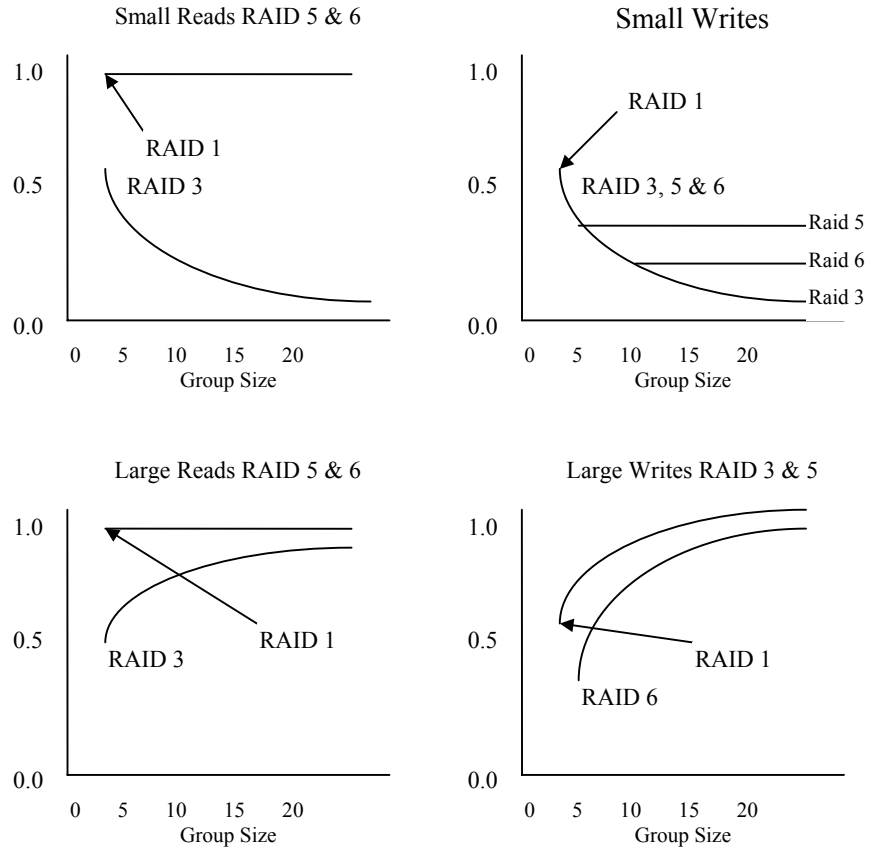
Ο παραπάνω πίνακας συγκρίνει τις ρυθμοαποδόσεις σε πέντε σχήματα πλεονασμού για τέσσερις τύπους αιτήσεων I/O. Οι αιτήσεις περιορισμένου όγκου, (small read/write), αντιστοιχούν σε ένα striping unit και αφορούν ένα δίσκο. Οι αιτήσεις μεγάλου όγκου, (large read/write), αφορούν όλους τους δίσκους της συστοιχίας. Η παράμετρος G αντιστοιχεί στον αριθμό των δίσκων που συγκροτούν μίας ομάδα για την οποία υπολογίζεται πληροφορία ισοτιμίας, (parity group size).

Το κόστος του κάθε συστήματος θεωρείται ανάλογο του μέγιστου αριθμού δίσκων της συστοιχίας. Ετσι, ο παραπάνω πίνακας δείχνει ότι στις συστοιχίες RAID 0 και RAID 1, ισοδύναμου κόστους, το RAID 1 μπορεί να υποστηρίξει τον μισό αριθμό εγγραφών περιορισμένου όγκου, (small writes), από αυτόν που μπορεί να υποστηρίξει η συστοιχία RAID 0. Ισοδύναμα, ισχύει ότι το κόστος των εγγραφών περιορισμένου όγκου, (small writes), στην συστοιχία RAID 1, είναι διπλάσιο από το αντίστοιχο του RAID 0.

Πέρα από τις επιδόσεις, ο παραπάνω πίνακας επιδεικνύει την αποδοτικότητα αποθήκευσης, (storage efficiency), των δεδομένων στην κάθε συστοιχία. Η αποδοτικότητα αποθήκευσης είναι, κατά προσέγγιση, αντίστροφη του κόστους της κάθε μονάδας της χωρητικότητας χρήστη σε σχέση με την αντίστοιχη τιμή της συστοιχίας RAID 0. Για τις παραπάνω οργανώσεις δίσκων, η αποδοτικότητα αποθήκευσης ισούται με την μετρική performance/cost για τις εγγραφές μεγάλου όγκου, (large writes).

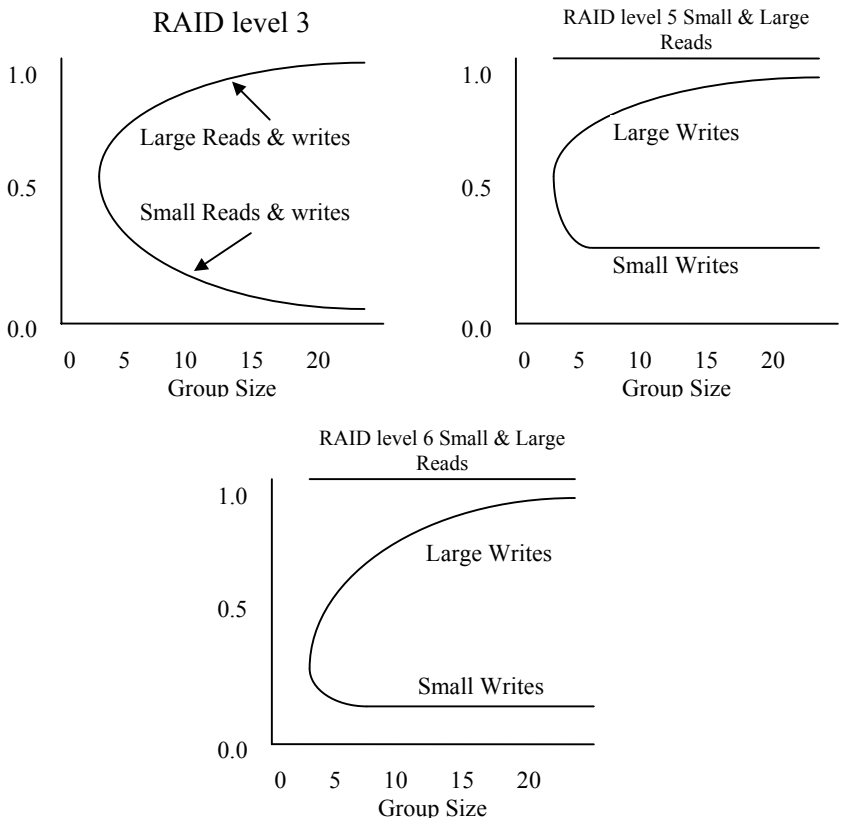
Στο Σχήμα 7 που ακολουθεί απεικονίζεται η μετρική επιδόσεων/κόστος του Πίνακα 2 για τα επίπεδα RAID 1, 3, 5 και 6, συναρτήσει του μεγέθους των ομάδων ισοτιμίας, (G). Παρατηρείται ότι η επίδοση/κόστος της συστοιχίας RAID 1 είναι ίση με αυτήν του συστήματος RAID 5, όταν το μέγεθος της ομάδας ισοτιμίας ισούται με

δύο, (G=2). Επίσης, η επίδοση/κόστος του επιπέδου RAID 3 είναι πάντοτε μικρότερη ή ίση από αυτήν του RAID 5. Στην πραγματικότητα όμως είναι δυνατόν να κατασκευαστεί ειδικό σύστημα RAID επιπέδου 3 το οποίο να είναι καλύτερο σε επίδοση/κόστος από μια αντίστοιχη ειδική υλοποίηση του συστήματος RAID 5. Επίσης στην περίπτωση των εγγραφών περιορισμένου όγκου, (small writes), τα συστήματα 3, 5 και 6 συμπεριφέρονται το ίδιο για μικρά G, ενώ καθώς το G αυξάνει τα RAID 5 και 6 πετυχαίνουν καλύτερη επίδοση.



Σχήμα 7: Ρυθμοαπόδοση/νομισματική μονάδα, (USD), συναρτήσει του G

Οπως επισημάνθηκε και παραπάνω, στο ερώτημα για το ποιο από τα 7 επίπεδα RAID θα ήταν το καταλληλότερο, η απάντηση θα ήταν ανάλογη του πλήθους δίσκων της ομάδας ισοτιμίας, (G), καθώς και του striping unit. Για παράδειγμα, αν G=2, τότε η καταλληλότερη συστοιχία θα ήταν αυτή που θα υιοθετούσε την τεχνική disk mirroring. Στην περίπτωση το striping unit ήταν πολύ μικρό το επίπεδο RAID 3 θα κρινόταν ικανοποιητικό. Για την υποστήριξη της λήψης ανάλογων αποφάσεων το Σχήμα 8 απεικονίζει τις τέσσερις μετρικές επίδοσης/κόστους του Πίνακα 2, για τις συστοιχίες 3, 5 και 6.



Σχήμα 7: Ρυθμοαπόδοση/νομισματική μονάδα για τις συστοιχίες 3, 5 και 6

4.3 ΑΞΙΟΠΙΣΤΙΑ

Ως μετρική σε συστήματα I/O, η αξιοπιστία είναι εξίσου σημαντική με την επίδοση, (performance), και το κόστος, (cost). Η μεγάλη αξιοπιστία που παρατηρείται στα συστήματα RAID τα καθιστά ιδιαίτερα δημοφιλή στην αγορά πληροφορικής. Στις παραγράφους που ακολουθούν αναλύεται η αξιοπιστία των συστοιχιών αυτών.

4.3.1 Βασική Αξιοπιστία

Ο πλεονασμός δεδομένων που παρατηρείται στις συστοιχίες δίσκων χρησιμοποιείται για την εξάλειψη της πιθανότητας καταστροφής της πληροφορίας από την δυστοκία, (failure), ορισμένων δίσκων. Στην περίπτωση που θεωρηθεί ότι υπάρχουν μόνο ανεξάρτητες μεταξύ τους βλάβες συσκευών, ένα απλό σχήμα ισοτιμίας θα κρινόταν επαρκές.

Στο [PAT88], υπολογίζεται ο **MTTF**, (**mean time to failure**), για την συστοιχία επιπέδου 5 ως εξής:

$$\frac{\mathrm{MTTF(disk)}^2}{\mathrm{N} \times (\mathrm{G}-1) \times \mathrm{MTTR(disk)}}$$

Οπου **MTTF(disk)** είναι ο MTTF ενός δίσκου. **MTTR(disk)** είναι ο μέσος χρόνος που παρέρχεται μέχρις ότου επιδιορθωθεί ένας προβληματικός δίσκος, (**mean time to repair**). **N** είναι ο μέγιστος αριθμός δίσκων στην συστοιχία, και **G** το μέγεθος της ομάδας ισοτιμίας. Αν υποτεθεί για παράδειγμα ότι είναι διαθέσιμοι N=100 δίσκοι, ο κάθε δίσκος εμφανίζει MTTF(disk)=200.000 ώρες και MTTR=1 ώρα. Εάν οι παραπάνω δίσκοι οργανωθούν σε ομάδες ισοτιμίας που κατά μέσο όρο έχουν μέγεθος G=16, τότε ο MTTF του συστήματος θα είναι, σύμφωνα με τον παραπάνω τύπο, 3000 χρόνια.

Για μία συστοιχία με δύο δίσκους πλεονασμού για κάθε ομάδα ισοτιμίας, όπως είναι το σχήμα P+Q, (επίπεδο 6), που αναφέρθηκε, ο MTTF υπολογίζεται ως εξής:

$$\frac{\mathrm{MTTF(disk)}^3}{\mathrm{N} \times (\mathrm{G}-1) \times (\mathrm{G}-2) \times \mathrm{MTTR}^2\mathrm{(disk)}}$$

Χρησιμοποιώντας και εδώ τις ίδιες τιμές, υπολογίζεται για τον MTTF η αστρονομική τιμή των 38 εκατομμυρίων χρόνων.

Αυτές βέβαια οι τιμές αποτελούν μια ιδεατή εικόνα, παρόλα αυτά όμως παρέχουν μία ιδέα για το μέγεθος της αξιοπιστίας που μπορούν να πετύχουν οι συστοιχίες δίσκων. Στην συνέχεια θα εξεταστεί η αξιοπιστία συστημάτων block interleaved στις συγκεκριμένες περιπτώσεις κατάρρευσης συστήματος, (system crash), μη διορθώσιμων λαθών, (uncorrectable bit errors), και συσχετιζόμενων βλαβών δίσκων, (correlated bit errors), που επηρεάζουν σε σημαντικό βαθμό την αξιοπιστία τέτοιων συστημάτων.

4.3.2 Καταρρεύσεις συστήματος και ασυνέπεια στην πληροφορία ισοτιμίας

Στην παράγραφο αυτή ο όρος κατάρρευση συστήματος, (system crash), αναφέρεται στα γεγονότα εκείνα που μπορεί να συμβούν, όπως βλάβη στην παροχή ισχύος, λάθος κατά την διαχείριση του συστήματος, βλάβη στο υλικό ή κάποιο λάθος λογισμικού και είναι δυνατόν να διακόψουν μια λειτουργία εισόδου/εξόδου σε μία συστοιχία.

Τέτοιου είδους καταρρεύσεις μπορεί να διακόψουν τις λειτουργίες εγγραφής, (write operations), έχοντας πολλές φορές σαν αποτέλεσμα τα κύρια δεδομένα να

έχουν ενημερωθεί ενώ τα αντίστοιχα της ισοτιμίας όχι, (ή αντίστροφα). Σε κάθε περίπτωση το πρόβλημα εντοπίζεται στο γεγονός ότι οι πληροφορίες ισοτιμίας δεν είναι συνεπείς με τα κύρια δεδομένα και έτσι δεν μπορούν να χρησιμοποιηθούν όταν εμφανίσει πρόβλημα κάποιος δίσκος. Εδώ εφαρμόζονται κάποιες τεχνικές πλεονασμού του υλικού ή των παροχών ισχύος έτσι ώστε να μειωθεί η συχνότητα εμφάνισης των καταρρεύσεων. Εντούτοις, θα πρέπει να επισημανθεί ότι το συγκεκριμένο πρόβλημα δεν μπορεί να εξαλειφθεί πλήρως, (σε ποσοστό 100%).

Οι καταρρεύσεις συστήματος δημιουργούν ασυνέπειες στην πληροφορία ισοτιμίας και στην περίπτωση των bit-interleaved αλλά και στην περίπτωση των block-interleaved συστοιχιών. Το πρόβλημα όμως έχει πρακτική σημασία μόνο στην περίπτωση των block-interleaved διότι στις bit-interleaved τα δεδομένα που επηρεάζονται είναι μόνο αυτά που ενημερώνονταν κατά την διάρκεια της κατάρρευσης. Σε μία block-interleaved συστοιχία η διακοπή της λειτουργίας εγγραφής μπορεί να επηρεάσει την πληροφορία ισοτιμίας πολλαπλών blocks δεδομένων.

Ουσιαστικά, οι καταρρεύσεις συστήματος θεωρούνται πιο επιβλαβείς από τις βλάβες δίσκων για δύο λόγους:

- Συμβαίνουν πιο συχνά από αυτές.
- Μια κατάρρευση σε ένα σύστημα P+Q είναι ισοδύναμη με την βλάβη δύο συσκευών δίσκων, εφόσον τα δεδομένα “P” αλλά και τα δεδομένα “Q” καθίστανται ασυνεπή.

Για να αποφευχθεί η απώλεια πληροφορίας ισοτιμίας κατά τις καταρρεύσεις, θα πρέπει να τοποθετηθεί σε κάποια μη πτητική μνήμη, (non-volatile storage), πριν εκτελεστεί η λειτουργία της εγγραφής, τόση πληροφορία όση χρειάζεται για την ανάκτηση της πληροφορίας ισοτιμίας που ενδεχομένως καταστραφεί. Η πληροφορία αυτή θα πρέπει να διατηρείται μέχρις ότου ολοκληρωθεί η λειτουργία εγγραφής. Ειδικοί σχεδιασμοί των συστημάτων RAID μπορούν να υποστηρίξουν τέτοιου είδους λειτουργίες χρησιμοποιώντας μη πτητικές RAM.

4.3.3 Μη διορθώσιμα λάθη πληροφορίας, (Uncorrectable bit errors)

Παρά το γεγονός ότι οι σημερινοί δίσκοι είναι συσκευές υψηλής αξιοπιστίας, πολλές φορές αποτυγχάνουν στην σωστή εγγραφή ή ανάγνωση bits πληροφορίας. Ο ρυθμός με τον οποίο συμβαίνουν τέτοιου είδους αποτυχίες για τους περισσότερους

σύγχρονους δίσκους κυμαίνεται στο 1 λάθος bit κατά το διάβασμα 10^{14} bits. Σύμφωνα με τους κατασκευαστές των δίσκων τα λάθη αυτά δημιουργούνται κατά τις λειτουργίες εγγραφής, (write operations), και ανιχνεύονται κατά τις λειτουργίες ανάγνωσης, (read operations). Ετσι, η παραπάνω τιμή $1/10^{14}$ bits εκφράζει τον ρυθμό ανίχνευσης τέτοιων λαθών κατά την ομαλή λειτουργία του δίσκου.

Εστω για παράδειγμα ότι η επανακατασκευή ενός προβληματικού δίσκου σε μία συστοιχία των 100 GB απαιτεί επιτυχή ανάγνωση πληροφορίας 200 εκατομμυρίων sectors. Ο ρυθμός 1 λάθος στα 10^{14} bits, υποδηλώνει ότι ένας sector μεγέθους 512 byte στους 24 δισεκατομμύρια sectors θα διαβαστεί λάθος. Ετσι, αν υποτεθεί ότι η πιθανότητα ανεπιτυχούς ανάγνωσης ενός sector, διαφέρει από δίσκο σε δίσκο, τότε η πιθανότητα σωστής ανάγνωσης 200 εκατομμυρίων sectors θα είναι κατά προσέγγιση:

$$(1 - 1 / (2.4 \times 10^{10})) \wedge (2.0 \times 10^{8}) = 99.2\%$$

Η τιμή που υπολογίστηκε παραπάνω υποδεικνύει ότι κατά μέσο όρο το 0.8% των αποτυχιών στους δίσκους που έχει σαν αποτέλεσμα την απώλεια δεδομένων, οφείλεται σε μη διορθώσιμα λάθη πληροφορίας. Ετσι για τους κατασκευαστές δίσκων αυτού του είδους τα λάθη αποτελούν έναν αξιοπρόσεκτο παράγοντα.

Μια προσέγγιση που χρησιμοποιείται για την προστασία από τα μη διορθώσιμα λάθη πληροφορίας αφορά στην παρακολούθηση και εποπτεία των προειδοποιητικών σημάτων, (warnings), που παράγονται από τους δίσκους. Εάν διαφαίνεται η πιθανότητα βλάβης της συσκευής θα πρέπει να ενεργοποιηθούν κατάλληλες διαδικασίες. Η προσέγγιση αυτή υιοθετείται στο σύστημα VAXsimPLUS της DEC.

4.3.4 Συσχετιζόμενες βλάβες δίσκων, (Correlated disk failures)

Το απλούστερο μοντέλο αξιοπιστίας στις συστοιχίες δίσκων [PAT88], υποθέτει ότι όλες οι βλάβες, (failures), είναι μεταξύ τους ανεξάρτητες, για τον υπολογισμό του μέσου χρόνου μέχρις ότου σημειωθούν απώλειες δεδομένων, (**mean time to data loss, MTTDL**). Αυτό έχει σαν αποτέλεσμα οι υπολογισμοί να οδηγούν σε αποτελέσματα της τάξης των εκατομμυρίων ετών. Στην πραγματικότητα όμως, πολλοί περιβαλλοντικοί και κατασκευαστικοί παράγοντες οδηγούν συχνά σε βλάβες δίσκων που σχετίζονται μεταξύ τους. Για παράδειγμα ένας σεισμός είναι δυνατόν να αυξήσει τον ρυθμό εμφάνισης των βλαβών για όλους τους δίσκους μιας συστοιχίας,

σε μία μικρή χρονική περίοδο. Παρόμοιες βλάβες μπορούν να προκληθούν από προβλήματα στην παροχή ισχύος ή στο υλικό, (H/W), που είναι κοινό σε όλη την συστοιχία.

Ανεξάρτητα όμως από τους περιβαλλοντικούς παράγοντες, οι δίσκοι είναι δυνατό να παρουσιάσουν εκ φύσεως συσχετιζόμενες βλάβες. Για παράδειγμα είναι πολύ πιθανό για έναν δίσκο να αποτύχει είτε στην αρχή, (infant mortality), είτε στο τέλος της ζωής του, (wearout). Οι αποτυχίες στην αρχή της ζωής του οφείλονται, κατά κύριο λόγο, στα παροδικά ελαττώματα, (temporary defects), που δεν ανιχνεύθηκαν από τους κατασκευαστές. Οι αποτυχίες στο τέλος της ζωής του οφείλονται στο ότι ο δίσκος έχει πια παλιώσει.

Οι συσχετιζόμενες αποτυχίες δίσκων ελαττώνουν σημαντικά την αξιοπιστία της συστοιχίας αφού μια βλάβη σε έναν δίσκο μπορεί να ακολουθείται από μια άλλη πριν ακόμη επιδιορθωθεί.

4.3.5 Στατιστικά αποτελέσματα

Στις παραγράφους που προηγήθηκαν αναλύθηκε πως οι καταρρεύσεις συστήματος, τα μη διορθώσιμα λάθη και οι συσχετιζόμενες βλάβες ελαττώνουν την αξιοπιστία των συστοιχιών. Στην συνέχεια θα υπολογιστεί ο μέσος χρόνος μέχρις ότου να υπάρξει απώλεια δεδομένων, (**mean time to data loss, MTTDL**), λαμβάνοντας υπόψη τους παράγοντες αυτούς.

Υπάρχουν τρεις κοινοί τρόποι για την απώλεια δεδομένων σε μία block-interleaved συστοιχία που εφαρμόζει μηχανισμούς ισοτιμίας:

- Διπλή βλάβη δίσκου.
- Κατάρρευση συστήματος ακολουθούμενη από βλάβη σε δίσκο.
- Βλάβη σε δίσκο ακολουθούμενη από ένα μη διορθώσιμο λάθος πληροφορίας κατά την διάρκεια της επανακατασκευής.

Οπως αναφέρθηκε παραπάνω η απώλεια που προέρχεται από την κατάρρευση συστήματος που ακολουθείται από βλάβη σε έναν δίσκο, είναι δυνατόν να αποφευχθεί με την χρήση της μη πτητικής μνήμης, (non-volatile storage), σε H/W υλοποιήσεις συστοιχιών. Ομως τέτοιου είδους προστασία είναι αδύνατη για τις συστοιχίες που υλοποιούνται με την χρήση ειδικού λογισμικού. Γενικότερα, οι παραπάνω τρεις περιπτώσεις συσχετιζόμενης αποτυχίας είναι και οι πιο δύσκολες για να αποφευχθούν.

Για να δημιουργηθεί ένα απλό μοντέλο συσχετιζόμενων βλαβών δίσκων, θα υποτεθεί ότι η κάθε αποτυχία που θα ακολουθεί θα είναι 10 φορές πιο πιθανό να συμβεί από την προηγούμενή της, (μέχρις ότου επανακατασκευαστεί ο προβληματικός δίσκος). Ο Πίνακας 3 που ακολουθεί παρουσιάζει τιμές των διαφόρων παραμέτρων αξιοπιστίας, οι οποίες θα χρησιμοποιηθούν για τον υπολογισμό αριθμητικών τιμών. Η χωρητικότητα που είναι διαθέσιμη στον χρήστη είναι 100 δίσκοι, (500 GB), ενώ 16 από αυτούς χρησιμοποιούνται για την αποθήκευση πληροφορίας ισοτιμίας.

Πίνακας 3: Παράμετροι αξιοπιστίας

Συνολική χωρητικότητα χρήστη, (user capacity)	100 δίσκοι (500 GB)
Μέγεθος δίσκου (disk size)	5 GB
Μέγεθος τομέα (sector size)	512 bytes
Ρυθμός εμφάνισης λαθών (bit error rate, BER)	1 σε 10^{14} bits ή 1 σε 2.4×10^{10} sectors
p(disk): Η πιθανότητα ανάγνωσης όλων των τομέων σε έναν δίσκο.	99.96 %
Μέγεθος ομάδας ισοτιμίας (parity group, G)	16 δίσκοι
MTTF(disk): για έναν δίσκο	200000 ώρες
MTTF(disk2): για δύο δίσκους	20000 ώρες
MTTF(disk3): για τρεις δίσκους	2000 ώρες
MTTR(disk): για έναν δίσκο	1 ώρα
MTTF(sys): για το σύστημα	1 μήνα
MTTR(sys): για το σύστημα	1 ώρα

Στον Πίνακα 4, στον οποίο παρουσιάζονται οι μετρικές αξιοπιστίας για την συστοιχία RAID 5, φαίνεται ότι οι συχνότητες με τις οποίες εμφανίζονται οι τρεις παραπάνω συσχετιζόμενες αποτυχίες, διαφέρουν η μία της άλλης κατά μία τάξη μεγέθους. Αυτό σημαίνει ότι για την μέτρηση της αξιοπιστίας θα πρέπει να ληφθούν υπόψη και ο τρεις αυτοί παράγοντες. Γίνεται έτσι δύσκολο να βελτιωθεί η ολική αξιοπιστία ενός συστήματος. Για παράδειγμα ένας πιο αξιόπιστος δίσκος θα πρέπει να έχει μειώσει σημαντικά την συχνότητα εμφάνισης των διπλών αποτυχιών. Επίσης τόσο οι πτώσεις συστήματος, όσο και η ρυθμοί εμφάνισης λαθών, θα πρέπει να ελαττωθούν ώστε να γίνουν οι βελτιώσεις στην αξιοπιστία. Ακόμη είναι προφανές πως παρά το ότι για τις διπλές αποτυχίες ο χρόνος MTTDL είναι 285 χρόνια, υπάρχει μία πιθανότητα 3.4% να χαθούν δεδομένα στα 10 πρώτα χρόνια.

Πίνακας 4: Χαρακτηριστικά βλαβών για συστήματα επιπέδου 5

Συσχετιζόμενες αποτυχίες	MTTDL	MTTDL	Pr.
α. Διπλή αποτυχία δίσκου	$\frac{MTTF(disk) \times MTTF(disk2)}{N \times (G-1) \times MTTR(disk)}$	285 χρόνια	3.4%
β. Κατάρρευση συστήματος + αποτυχία δίσκου	$\frac{MTTF(sys) \times MTTF(disk)}{N \times MTTR(sys)}$	154 χρόνια	6.3%
γ. Αποτυχία δίσκου + μη επιδιορθώσιμα λάθη	$\frac{MTTF(disk)}{N \times (1-(p(disk))^{G-1})}$	36 χρόνια	24.4%
Software RAID	(Αρμονικό άθροισμα των α, β, γ)	26 χρόνια	31.6%
Hardware RAID (NVRAM)	(Αρμονικό άθροισμα, των α, γ)	32 χρόνια	26.8%

Οπου Pr, η πιθανότητα να απώλειας δεδομένων στην περίοδο των 10 χρόνων. N ισούται με τον αριθμό των δίσκών (100) επί G/(G-1).

Στον Πίνακα 5, απεικονίζονται οι μετρικές αξιοπιστίας για την συστοιχία P+Q, (επίπεδο 6). Θα πρέπει να τονιστεί ότι οι καταρρεύσεις αποτελούν το ευαίσθητο σημείο αυτών των συστημάτων, εφόσον κατά την εμφάνιση τους οι πληροφορίες P και Q χάνουν την συνέπεια τους. Για τον λόγο αυτό στις συστοιχίες P+Q οι καταρρεύσεις θεωρούνται ισάξιες με τις διπλές αποτυχίες δίσκων. Ετσι εάν δεν έχουν ληφθεί μέτρα για την προστασία από τις καταρρεύσεις συστήματος, η χρήση της συστοιχίας P+Q δεν είναι καθόλου συμφέρουσα. Αντίθετα, η χρήση της θα πλεονεκτούσε σε περιπτώσεις που συνέβαιναν μη επιδιορθώσιμα λάθη κατά τις ανακατασκευές ή πολλαπλές αποτυχίες δίσκων.

Πίνακας 5: Χαρακτηριστικά βλαβών για συστοιχίες P+Q

Συσχετιζόμενες αποτυχίες	MTTDL	MTTDL	Pr.
α. Τριπλή αποτυχία δίσκου	$\frac{MTTF(disk) \times MTTF(disk2) \times MTTF(disk3)}{N \times (G-1) \times (G-2) \times MTTR^2(disk)}$	38052 χρόνια	0.03%
β. Κατάρρευση συστήματος + αποτυχία δίσκου	$\frac{MTTF(sys) \times MTTF(disk)}{N \times MTTR(sys)}$	144 χρόνια	7.7%
γ. Διπλή αποτυχία δίσκου + μη επιδιορθώσιμα λάθη	$\frac{MTTF(disk) \times MTTF(disk2)}{N \times (G-1) \times (1-(1-p(disk)))^{G-2} \times MTTR(disk)}$	47697 χρόνια	0.02%
Software RAID	(Αρμονικό άθροισμα των α, β, γ)	143 χρ.	6.8%
Hardware RAID, (NVRAM)	(Αρμονικό άθροισμα, των α, γ)	21166 χρ.	0.05%

Οπου Pr, η πιθανότητα απώλειας δεδομένων στην περίοδο των 10 χρόνων. N ισούται με τον αριθμό των δίσκών (100) επί G/(G-2).

4.3.6 Συμπεράσματα σχετικά με την αξιοπιστία

Από το περιεχόμενο των προηγούμενων παραγράφων θα πρέπει να τονιστούν τα εξής:

- Τα λάθη που προέρχονται από τις αποτυχίες που σχετίζονται μεταξύ τους παρουσιάζουν και την μεγαλύτερη δυσκολία στο να αποφευχθούν
- Η καταρρεύσεις συστήματος καθώς και τα μη επιδιορθώσιμα λάθη ελαττώνουν σημαντικά την αξιοπιστία των block interleaved συστοιχιών.
- Τα συστήματα P+Q είναι πολύ αποτελεσματικά για την προστασία από τις διπλές αποτυχίες στους δίσκους καθώς και από τα μη επιδιορθώσιμα λάθη που εμφανίζονται κατά τις ανακατασκευές, ενώ αντίθετα παρουσιάζουν μεγάλη ευαισθησία στις καταρρεύσεις συστήματος.
- Θα πρέπει στις P+Q συστοιχίες να χρησιμοποιηθούν μηχανισμοί προστασίας με μη μεταβλητές μνήμες (non-volatile storage), ώστε να αποφευχθούν οι καταρρεύσεις και να πετύχουμε την μεγαλύτερη δυνατή αξιοπιστία.

5. ΖΗΤΗΜΑΤΑ ΥΛΟΠΟΙΗΣΗΣ ΣΥΣΤΟΙΧΙΩΝ

Αν και η λειτουργία μίας block interleaved συστοιχίας δίσκων είναι σχετικά απλή, θα πρέπει, κατά την κατασκευή της, να ληφθούν υπόψη ορισμένα στοιχεία έτσι ώστε το όλο σύστημα να λειτουργεί σωστά, αξιόπιστα και με συγκεκριμένο επίπεδο

επιδόσεων. Ενα πρόβλημα, που εντοπίζεται, είναι ότι η πληροφορία κατάστασης, (status information), μίας συστοιχίας δίσκων δεν συντίθεται μόνο από τα δεδομένα και την πληροφορία ισοτιμίας. Πέρα των παραπάνω στοιχείων, υπάρχει και άλλη πληροφορία που σχετίζεται με το ποιος δίσκος έχει υποστεί βλάβη, το κατά πόσο ένας προβληματικός δίσκος έχει επανακατασκευαστεί και το ποιοι sectors πρόκειται να ενημερωθούν άμεσα.

Ολες αυτές οι συμπληρωματικές πληροφορίες θα πρέπει να φυλάσσονται με ακρίβεια, ώστε να χρησιμοποιηθούν όταν υπάρξει κάποια κατάρρευση του συστήματος. Για την πληροφορία αυτή χρησιμοποιείται ο όρος πληροφορία μετά-κατάστασης, (**metastate information**). Επίσης ένα άλλο πρόβλημα, το οποίο θίγεται στη παράγραφο 5.4, είναι ότι πολλαπλοί δίσκοι είναι συνήθως συνδεδεμένοι με την κεντρική μονάδα, (host computer), μέσω μίας κοινής αρτηρίας ή ενός καλωδίου.

5.1 ΑΠΟΦΥΓΗ ΠΡΟΒΛΗΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Το μόνο τμήμα της πληροφορίας μετά-κατάστασης που θα πρέπει οπωσδήποτε να φυλαχθεί σε μία συστοιχία δίσκων είναι το αν τα δεδομένα του κάθε τομέα, είναι έγκυρα (valid) ή όχι. Για να διατηρηθεί αυτή η πληροφορία θα πρέπει να εφαρμοστούν οι παρακάτω περιορισμοί:

- Μόλις υποστεί βλάβη ένας δίσκος, οι λογικοί sectors που αντιστοιχούν σε αυτόν θα πρέπει να σημαδευτούν ως μη έγκυροι, (**invalid**), πριν υπάρξει αίτηση από κάποιον χρήστη, για εγγραφή ή ανάγνωση δεδομένων από τους συγκεκριμένους τομείς. Η ενέργεια αυτή προφυλάσσει τους χρήστες από την ανάγνωση κατεστραμμένων δεδομένων από τις προβληματικές συσκευές.
- Μόλις ένας μη έγκυρος λογικός τομέας επανακατασκευαστεί, σε μία εφεδρική συσκευή, θα πρέπει να σημαδευτεί ως έγκυρος, (valid), πριν διεκπεραιωθεί αίτηση εγγραφής που φυσιολογικά θα έγραφε σε έναν προβληματικό δίσκο. Η αλλαγή αυτή εξασφαλίζει ότι ακολουθιακές εγγραφές ενημερώνουν τα επανακατασκευάσιμα δεδομένα στον εφεδρικό δίσκο.

Οι παραπάνω περιορισμοί αποτρέπουν την λήψη, από τις συσκευές δίσκων παρωχημένων δεδομένων. Λόγω του γεγονότος ότι οι βλάβες δίσκων εμφανίζονται αρκετά σπάνια, και μεγάλες ομάδες stripping units είναι δυνατόν κάποια στιγμή να καταστούν μη έγκυρες, η ενημέρωση της πληροφορίας μετακατάτασης σε χώρο ασφαλούς περιφερειακής αποθήκευσης, (stable storage), δεν έχει σημαντικές επιπτώσεις στην επίδοση.

5.2 ΑΝΑΚΑΤΑΣΚΕΥΗ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΙΣΟΤΙΜΙΑΣ ΜΕΤΑ ΑΠΟ ΚΑΤΑΡΡΕΥΣΗ ΣΥΣΤΗΜΑΤΟΣ

Οι καταρρεύσεις συστήματος έχουν σαν αποτέλεσμα να δημιουργούνται ασυνέπειες στην πληροφορία ισοτιμίας όταν διακόπτονται λειτουργίες εγγραφής. Ετσι, εκτός κι αν είναι γνωστός ο parity sector που ενημερωνόταν κατά την στιγμή της κατάρρευσης, θα πρέπει όλοι οι parity sectors να ανακατασκευαστούν μόλις επανέλθει η συστοιχία. Αυτή είναι μια ιδιαίτερα ακριβή διαδικασία εφόσον απαιτεί ανίχνευση των περιεχομένων όλης της συστοιχίας. Για να αποφευχθεί το σχετικό κόστος, θα πρέπει να φυλάσσεται κατάλληλη πληροφορία, (στον δίσκο), που να διευκρινίζει την συνεπή ή μη κατάσταση του κάθε τομέα. Οι παρακάτω, λοιπόν, περιορισμοί θα πρέπει να ληφθούν υπόψη:

- Πριν την εξυπηρέτηση οποιασδήποτε αίτησης εγγραφής, οι αντίστοιχοι τομείς ισοτιμίας, (parity sectors), θα πρέπει να χαρακτηριστούν ως ασυνεπείς.
- Οταν το σύστημα επανέλθει μετά από μία κατάρρευση, όλοι οι τομείς που είναι ασυνεπείς θα πρέπει να ανακατασκευαστούν.
- Η ανακατασκευή της πληροφορίας ισοτιμίας δεν προκαλεί δυσλειτουργίες στον δίσκο. Για τον λόγο αυτό, συχνά ορισμένοι τομείς χαρακτηρίζονται ως ασυνεπής παρά το γεγονός ότι δεν έχουν παρουσιάσει πρόβλημα, απλά και μόνο για να υπάρξει μία ανακατασκευή της πληροφορίας ισοτιμίας, Για να αποφευχθεί όμως η αναπαραγωγή μεγάλου αριθμού τομέων ισοτιμίας, θα πρέπει σε τακτά χρονικά διαστήματα αυτοί να χαρακτηρίζονται ως συνεπείς. Διαφορετικά είναι προτιμητέο, μετά από κάθε λειτουργία εγγραφής, να χαρακτηρίζεται συνεπής ο αντίστοιχος τομέα ισοτιμίας.

5.3 ΛΕΙΤΟΥΡΓΙΑ ΜΕ ΥΠΑΡΞΗ ΠΡΟΒΛΗΜΑΤΙΚΟΥ ΔΙΣΚΟΥ

Η κατάρρευση ενός συστήματος σε μία block-interleaved συστοιχία δίσκων είναι παρόμοια με την αποτυχία ενός δίσκου, στο ότι και στις δύο περιπτώσεις υπάρχει απώλεια στην πληροφορία ισοτιμίας. Αυτό σημαίνει ότι σε μία συστοιχία που λειτουργεί με έναν προβληματικό δίσκο, είναι πολύ πιθανή η απώλεια δεδομένων με τη κατάρρευση του συστήματος. Ετσι, λόγω του ότι οι καταρρεύσεις έχουν μεγαλύτερη συχνότητα εμφάνισης από τις βλάβες των συσκευών δίσκων, το να λειτουργεί η συστοιχία με έναν προβληματικό δίσκο κρίνεται ιδιαίτερα ριψοκίνδυνο.

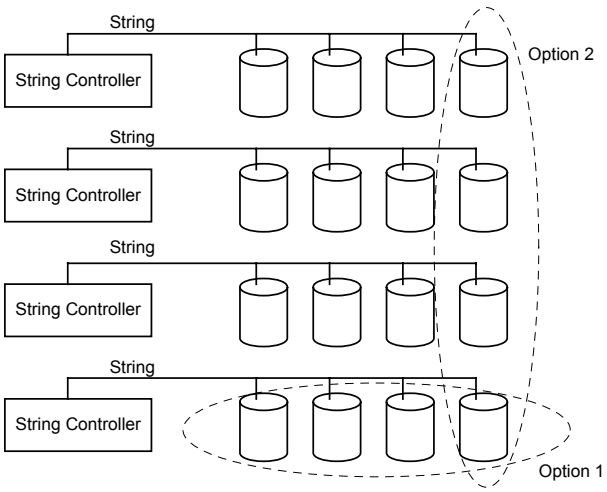
Στην περίπτωση που η συστοιχία λειτουργεί με ένα προβληματικό δίσκο, για να αποφευχθούν οι συνέπειες μιας ενδεχόμενης κατάρρευσης, εκτελείται πριν από κάθε λειτουργία εγγραφής ένα είδος προεγγραφής σε κάποια log περιοχή. Θα πρέπει να αναφερθεί ότι για να γίνει αυτή η προεγγραφή εφαρμόζονται συνήθως οι μέθοδοι:

- demand reconstruction ή
- parity sparing.

5.4 ΟΡΘΟΓΩΝΙΕΣ ΣΥΣΤΟΙΧΙΕΣ

Στην παράγραφο αυτή εξετάζεται το πρόβλημα της σύνδεσης των δίσκων της συστοιχίας στην κεντρική μονάδα. Η επιλογή του τρόπου σύνδεσης επηρεάζει σημαντικά την επίδοση και την αξιοπιστία του συστήματος. Οι περισσότεροι υπολογιστές συνδέουν πολλαπλούς δίσκους μέσω ενός μικρότερου αριθμού καλωδίων, (**strings**). Ετσι, μία βλάβη σε ένα καλώδιο προκαλεί ταυτόχρονες βλάβες στους συνδεόμενους με αυτό δίσκους, (single point of failure).

Σαν παράδειγμα, θα εξεταστεί η συστοιχία των 16 δίσκων του σχήματος 7 καθώς και οι δύο επιλογές για την οργάνωση ομάδων ισοτιμίας.



Σχήμα 9: Ορθογώνιο RAID

Η επιλογή 1 συνδυάζει κάθε καλώδιο με τέσσερις δίσκους σε μία ομάδα ισοτιμίας. Η επιλογή 2 συνδυάζει έναν δίσκο από κάθε καλώδιο σε μία ομάδα ισοτιμίας. Είναι φανερό ότι αν για την πρώτη περίπτωση καταστραφεί το καλώδιο, όλοι οι δίσκοι που είναι συνδεδεμένοι με αυτό και αποτελούν ομάδα θα είναι αδύνατο

να προσπελαθούν. Από την άλλη μεριά, στην επιλογή 2, χάνεται μόνο ένας δίσκος από την ομάδα, ενώ υπόλοιποι παραμένουν διαθέσιμοι.

Αυτή η τεχνική ορθογώνιας οργάνωσης των ομάδων ισοτιμίας, (διόρθωσης λαθών), ονομάζεται **ορθογώνιο RAID**. Η οργάνωση αυτή έχει το πλεονέκτημα να ελαχιστοποιεί τις συγκρούσεις, (conflicts), στα καλώδια, όταν πολλοί δίσκοι από διαφορετικές ομάδες μεταφέρουν ταυτόχρονα δεδομένα.

6. ΝΕΕΣ ΕΞΕΛΙΞΕΙΣ - ΒΕΛΤΙΩΣΕΙΣ

Στις παραγράφους που ακολουθούν παρουσιάζονται ορισμένες νέες τεχνικές για την διαμόρφωση και λειτουργία συστοιχιών δίσκων.

6.1 ΒΕΛΤΙΩΣΗ ΤΗΣ ΕΠΙΔΟΣΗΣ ΤΟΥ ΕΠΙΠΕΔΟΥ 5 ΣΕ ΕΓΓΡΑΦΕΣ ΠΕΡΙΟΡΙΣΜΕΝΟΥ ΟΓΚΟΥ

Το βασικότερο πρόβλημα του επιπέδου 5 είναι η υψηλή χρονική επιβάρυνση, (overhead), για εγγραφές περιορισμένου όγκου, (small writes). Οπως αναφέρθηκε στην παράγραφο 3.6 η λειτουργία εγγραφής περιορισμένου όγκου δεδομένων προϋποθέτει 4 I/O λειτουργίες, (δύο για να διαβαστούν τα παλιά δεδομένα και η παλιά πληροφορία ισοτιμίας και δύο για να γραφούν τα νέα δεδομένα και η νέα πληροφορία ισοτιμίας). Στις παραγράφους που ακολουθούν παρουσιάζονται ορισμένες στρατηγικές για την αντιμετώπιση του προβλήματος.

Buffering και caching

Οι τεχνικές buffering και caching είναι άμεσα εφαρμόσιμες στις συστοιχίες δίσκων. Η τεχνική write buffering, η οποία επίσης ονομάζεται και asynchronous write, θεωρεί ότι μία λειτουργία εγγραφής έχει ολοκληρωθεί, (στέλνει την σχετική επιβεβαίωση στο λειτουργικό ή στην εφαρμογή), πριν ακόμη τα δεδομένα καταλήξουν στον δίσκο. Ετσι, ελαττώνεται ο χρόνος απόκρισης για τον χρήστη κάτω από συνθήκες μέτριου και μικρού φόρτου. Εφόσον ο χρόνος απόκρισης δεν εξαρτάται πλέον από την συσκευή, η τεχνική buffering μπορεί να επιφέρει ανάλογα αποτελέσματα και στην περίπτωση που εφαρμοστεί σε συστοιχίες επιπέδου 5. Για να αποφευχθεί η απώλεια δεδομένων σε ενδεχόμενες καταρρεύσεις θα πρέπει να χρησιμοποιηθούν μη-πτητικές μνήμες, (non-volatile memories). Η τεχνική μπορεί να επιφέρει βελτίωση στην ρυθμοαπόδοσης του υποσυστήματος δίσκου κατά δύο τρόπους: (1) όταν υπάρχουν διαδοχικές ενημερώσεις των ιδίων δεδομένων, μόνο η

νεότερη μεταφέρεται στον δίσκο ενώ η παλαιότερη απορρίπτεται από τον χώρο του buffer, (2) η ουρά αιτήσεων επιμηκύνεται δίνοντας έτσι την ευκαιρία στον δρομολογητή δίσκου, (disk scheduler), να κάνει βελτιστοποιήσεις κατά την εκτέλεση τους.

Σε καταστάσεις υψηλού φόρτου, ο buffer εγγραφής θα υπερχειλίσει πολύ σύντομα και ο χρόνος απόκρισης στο επίπεδο 5 θα είναι 4 φορές μεγαλύτερος από αυτόν του επιπέδου 0. Μία άλλη προσέγγιση η οποία ακολουθείται, με ιδιαίτερα θετικά αποτελέσματα στην περίπτωση συστοιχιών επιπέδου 5, είναι η ομαδοποίηση των σειριακών εγγραφών, (grouping of sequential writes). Με την ομαδοποίηση, οι εγγραφές περιορισμένου όγκου, οι οποίες αποτελούν και το βασικό πρόβλημα στο επίπεδο 5, μπορούν να διαμορφώσουν εγγραφές που να καλύπτουν πλήρη stipping units.

Το read caching στο επίπεδο 5 χρησιμοποιείται και για τον υπολογισμό της νέας τιμής ισοτιμίας, στην περίπτωση που τα παλαιότερα σχετικά δεδομένα βρίσκονται ακόμη μέσα στην κρυφή μνήμη, (cache). Ετσι οι απαιτούμενες λειτουργίες I/O για το επίπεδο 5 ελαττώνονται από 4 σε 3. Η μέθοδος αυτή μπορεί να βρει άμεση εφαρμογή σε συστήματα επεξεργασίας δοσοληψιών, όπου οι εγγραφές ενημερώνονται ιδιαίτερα συχνά. Για την ενημέρωση μίας εγγραφής απαιτείται η ανάγνωση της παλιάς τιμής, η μεταβολή της και τέλος η καταχώρηση της νέας τιμής στην ίδια ακριβώς θέση.

Floating Parity

Για τον περιορισμό του χρονικού κόστους της διαδικασίας read-modify-write χρησιμοποιείται μία παραλλαγή της τυπικής οργάνωσης της πληροφορίας ισοτιμίας. Η παραλλαγή αυτή ονομάζεται floating parity. Η πληροφορία ισοτιμίας αποθηκεύεται σε κυλίνδρους που περιέχουν μία τουλάχιστον άτρακτο με κενά blocks. Οταν υπάρχει ανάγκη ενημέρωσης ενός parity block, το νέο parity block μπορεί να τοποθετηθεί στο κοντινότερο, κατά την περιστροφή, ελεύθερο block που έπεται του παλιού parity block. Για συσκευές με 16 ατράκτους ανά κύλινδρο, το πλησιέστερο ελεύθερο block ακολουθεί το parity block που μόλις διαβάστηκε, με πιθανότητα 0,65. Ο μέσος αριθμός των blocks που πρέπει να περάσουν από την κεφαλή πριν εντοπιστεί κενό block βρίσκεται μεταξύ 0,7 και 0,8. Με τον τρόπο αυτό η εγγραφή του νέου block ισοτιμίας μπορεί να πραγματοποιηθεί αμέσως μετά την ανάγνωση του παλιού, σε χρόνο που είναι κατά ένα millisecond μεγαλύτερος από τον χρόνο ανάγνωσης.

Parity Logging

Η τεχνική parity logging προσπαθεί να ελαχιστοποιήσει το χρονικό κόστος ενημέρωσης της πληροφορίας ισοτιμίας, καθυστερώντας τις σχετικές με αυτήν διαδικασίες ανάγνωσης και εγγραφής, (parity read & write). Αντί να ενημερωθεί άμεσα η πληροφορία ισοτιμίας, μία **update image**, η οποία εκφράζει την διαφορά μεταξύ της παλιάς και της καινούργιας ισοτιμίας, κρατείται προσωρινά σε ένα log. Η καθυστέρηση της ενημέρωσης επιτρέπει την ομαδοποίηση της πληροφορίας ισοτιμίας σε μεγάλα blocks τα οποία μπορούν να εγγραφούν πολύ πιο αποδοτικά.

Το **parity log** αποθηκεύεται προσωρινά σε μη-πτητική μνήμη. Οταν αυτή η μνήμη, η οποία έχει μέγεθος ορισμένων δεκάδων KB, γεμίσει η update image αποθηκεύεται σε μία ειδική περιοχή του δίσκου, (**log region**). Οταν και η περιοχή log του δίσκου γεμίσει, η update image φορτώνεται στην μνήμη και προστίθεται στην παλαιότερη πληροφορία ισοτιμίας. Το αποτέλεσμα αποτελεί την νέα πληροφορία ισοτιμίας. Οι μετακινήσεις δεδομένων από και προς τον δίσκο γίνονται σε μεγάλα blocks οπότε υπάρχει μία αποδοτικότερη χρήση του συστήματος.

6.2 DECLUSTERED PARITY

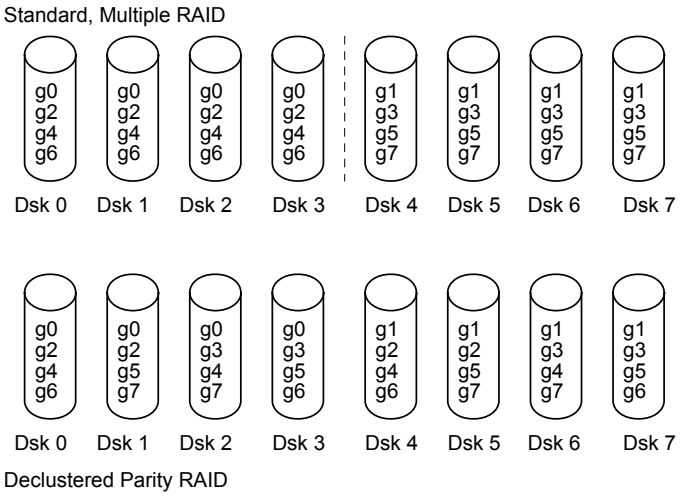
Πολλά από τα σύγχρονα συστήματα που βασίζονται σε συστοιχίες δίσκων απαιτούν από αυτές υψηλές επιδόσεις σε συνδυασμό με την αυξημένη διαθεσιμότητα των περιεχόμενων δεδομένων. Η απαίτηση αυτή εξακολουθεί να υφίσταται, για ορισμένες εφαρμογές, ακόμη και στην περίπτωση που συσκευές της συστοιχίας βρίσκονται εκτός λειτουργίας λόγω βλάβης και το σύστημα έχει εκκινήσει διαδικασία ανακατασκευής, (drive rebuild).

Σε συμβατικές συστοιχίες επιπέδου 5, οι βλάβες δίσκων υποβαθμίζουν σημαντικά τις επιδόσεις του συστήματος. Πιθανές αναφορές στα περιεχόμενα της προβληματικής συσκευής επιβάλλουν σημαντικά μεγαλύτερο αριθμό λειτουργιών I/O από τον συνήθη. Η αύξηση αυτή οφείλεται στην ανάγκη ανακατασκευής των δεδομένων.

Σε συστήματα όπου τα δεδομένα κατανέμονται σε περισσότερα του ενός RAIDs, (**interRAID stripping**), η μέση αύξηση του φόρτου είναι σημαντικά μικρότερη αλλά δεν παύει να υφίσταται στο parity group στο οποίο συνέβη η βλάβη. Το συγκεκριμένο group διαμορφώνει μία στενωπό, (bottleneck), και υποβαθμίζει την λειτουργία όλου του υποσυστήματος αποθήκευσης. Το πρόβλημα αφορά στην αδυναμία της συστοιχίας να κατανείμει ισόρροπα τον φόρτο εργασίας σε περίπτωση βλαβών. Μία τακτική για την αντιμετώπιση του προβλήματος είναι η parity

declustering. Θα πρέπει να τονιστεί ότι η τακτική αφορά υποσυστήματα περιφερειακής αποθήκευσης που συνθέτονται από περισσότερα του ενός RAIDs.

Στο Σχήμα 10 που ακολουθεί παρουσιάζονται οι διαφορές της parity declustering από την κλασσική προσέγγιση στην διαμόρφωση πολλαπλών συστοιχιών.



Σχήμα 10: Τεχνική Declustered Parity

Για την διαμόρφωση των συστοιχιών του σχήματος χρησιμοποιήθηκαν 8 συσκευές δίσκων ενώ το parity group size, (αριθμός δίσκων για τους οποίους υπολογίζεται πληροφορία ισοτιμίας), ήταν 4. Στην πρώτη περίπτωση διαμορφώνονται δύο συστοιχίες 4 δίσκων στις οποίες κατανέμονται τα δεδομένα ενώ τα parity groups δεν επικαλύπτονται. Το declustered parity RAID διαμορφώνεται από την επικάλυψη των parity groups, όπως αυτή φαίνεται στο δεύτερο σκέλος του σχήματος.

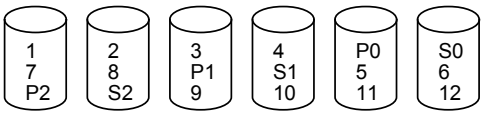
Εάν, για παράδειγμα, υποστεί βλάβη η συσκευή Dsk 2, η αίτηση για τα δεδομένα της θα προκαλέσει μεμονωμένες προσπελάσεις στις συσκευές Dsk 0, Dsk 1 και Dsk 3, στην περίπτωση που υιοθετηθεί η πρώτη οργάνωση, (Multiple RAIDs), ενώ δεν θα χρησιμοποιηθούν καθόλου οι συσκευές Dsk 4, Dsk 5, Dsk 6 και Dsk 7. Στην περίπτωση declustered parity RAID, η ίδια βλάβη θα προκαλέσει προσπελάσεις στις συσκευές Dsk 0, 1 και 3 για την αποκατάσταση του g0, στις συσκευές Dsk 3, 6 και 7 για την αποκατάσταση του g3, στις συσκευές Dsk 1, 4 και 6 για την αποκατάσταση του g4 και τέλος στις συσκευές Dsk 1, 5 και 6 για την αποκατάσταση του g7.

Από τα παραπάνω είναι προφανής η πληρέστερη κατανομή του φόρτου εργασίας η οποία επιτυγχάνεται, μέσω της declustered parity RAID.

6.3 ΕΚΜΕΤΑΛΛΕΥΣΗ ΤΩΝ ΕΦΕΔΡΙΚΩΝ ON-LINE ΣΥΣΚΕΥΩΝ

Στην παράγραφο αυτή περιγράφονται δύο τεχνικές για την εκμετάλλευση των εφεδρικών on-line δίσκων οι οποίοι παραμένουν πλήρως αδρανείς στην συστοιχία για ένα πολύ μεγάλο χρονικό διάστημα. Στόχο και των δύο τεχνικών αποτελεί η βελτίωση των επιδόσεων της συστοιχίας. Τα αποτελέσματα τους είναι περισσότερο αισθητά σε συστοιχίες με μικρό πλήθος δίσκων, επειδή το πηλίκο του πλήθους των εφεδρικών δίσκων ως το πλήθος των συνολικών είναι πολύ πιθανό να είναι μεγάλο.

Η τεχνική **distributed sparing** κατανέμει την χωρητικότητα των εφεδρικών συσκευών σε όλη την συστοιχία. Μία γραφική παρουσίαση της κατανομής παρέχει το Σχήμα 11 που ακολουθεί:

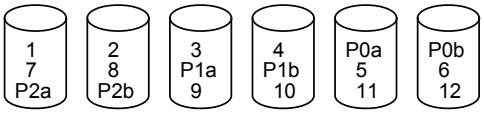


Σχήμα 11: Τεχνική Distributed Sparing

Η κατανομή της χωρητικότητας του εφεδρικού δίσκου είναι τελείως ανάλογη με αυτή της πληροφορίας ισοτιμίας που προβλέπεται από το επίπεδο RAID 5. Αντί να γίνεται χρήση N συσκευών και μία να παραμένει πάντα εφεδρική, γίνεται χρήση N+1 συσκευών, σε καθεμία από τις οποίες υπάρχει εφεδρικό τμήμα με χωρητικότητα ίση με το 1/(N+1) της χωρητικότητας της εφεδρικής συσκευής. Οταν μία συσκευή υποστεί βλάβη, τα blocks της ανακατασκευάζονται στα αντίστοιχα εφεδρικά blocks.

Η τεχνική επιτρέπει σε όλες τις συσκευές να συμμετάσχουν στην εξυπηρέτηση των αιτήσεων και, κατά συνέπεια, αυξάνει τις επιδόσεις της συστοιχίας. Ενα άλλο πλεονέκτημα της συγκεκριμένης τεχνικής είναι ότι, εφόσον οι δίσκοι δεν είναι πλήρεις, απαιτείται λιγότερη εργασία για την ανακατασκευή τους σε περίπτωση βλάβης.

Η τεχνική **parity sparing** είναι η ίδια με την distributed parity με την διαφορά ότι η εφεδρική χωρητικότητα χρησιμοποιείται για την αποθήκευση πληροφορίας ισοτιμίας. Μία γραφική παρουσίαση της κατανομής των δεδομένων ισοτιμίας δίνεται στο Σχήμα 12 που ακολουθεί:



Σχήμα 12: Τεχνική Parity Sparing

Οπως και στην περίπτωση distributed sparing η χρησιμοποίηση των εφεδρικών συσκευών βελτιώνει τις επιδόσεις της συστοιχίας. Με την εφαρμογή της parity sparing διπλασιάζονται τα blocks στα οποία αποθηκεύεται πληροφορία ισοτιμίας. Το νέο σύνολο των parity blocks, που προέκυψε, μπορεί να χρησιμοποιηθεί για τον λογικό διαμερισμό του συστήματος σε δύο συστοιχίες, πετυχαίνοντας υψηλότερη αξιοπιστία. Επίσης, το νέο σύνολο μπορεί να χρησιμοποιηθεί για την επέκταση των parity groups. Για την συστοιχία του σχήματος, εάν υποτεθεί ότι η πληροφορία ισοτιμίας των blocks 1, 2, 3, 4, P0a και P0b είναι πάντα 0, σε ενδεχόμενες εγγραφές μπορούν να ενημερωθούν είτε το P0a ή το P0b. Ετσι, παρέχεται στο σύστημα η δυνατότητα αποδοτικότερης δρομολόγησης των αιτήσεων εγγραφής ανάλογα με την κατάσταση της κάθε συσκευής δίσκου. Τέλος, το νέο σύνολο parity blocks μπορεί να χρησιμοποιηθεί για να αναπτυχθούν κώδικες Reed-Solomon, (P+Q).

Θα πρέπει να τονιστεί ότι οι παραπάνω τεχνικές βελτίωσης έχουν μελετηθεί μόνο μεμονωμένα και σε σημαντικά απλοποιημένα μοντέλα. Η μεταφορά τους σε εμπορικά προϊόντα απαιτεί σημαντική εργασία προς την κατεύθυνση της επίλυσης προβλημάτων συμβατότητας με άλλες τεχνολογίες, βελτιστοποίησης κ.α.

7. ΠΡΟΪΟΝΤΑ ΤΕΧΝΟΛΟΓΙΑΣ RAID

7.1 ΣΥΓΚΡΙΣΗ 4 ΣΥΣΤΟΙΧΙΩΝ ΔΙΣΚΩΝ

Στις παραγράφους που ακολουθούν παρουσιάζεται μία συγκριτική αξιολόγηση 4 συστημάτων RAID επιπέδου 5, [LAN94]. Τα εμπορικά διαθέσιμα προϊόντα των οποίων οι επιδόσεις εξετάζονται, είναι τα **Compaq Proliant**, **IBM Model 95**, **Micropolis Raidion** και **Conner RAID**.

Το σύστημα Compaq είχε την καλύτερη συμπεριφορά από τα υπόλοιπα 3 σε θέματα ρυθμοαπόδοσης και χρόνου προσπέλασης, απαιτούσε όμως σημαντικό χρόνο η ανακατασκευή δίσκου, (drive rebuild). Ο χρόνος ανακατασκευής δίσκου σε συστοιχίες RAID είναι το χρονικό εκείνο παράθυρο κατά την διάρκεια του οποίου το σύστημα είναι τρωτό, (**window of vulnerability**), εφόσον μία νέα βλάβη θα ήταν καταστροφική. Εξίσου ικανοποιητική, κρίθηκε η συμπεριφορά του Micropolis Raidion ενώ τις κατώτερες επιδόσεις εμφανίζουν το Model 95 και τέλος, το Conner. Τα συστήματα Compaq και IBM αποτελούν ενσωματωμένα συστατικά

ολοκληρωμένων μονάδων Η/Υ, (το array είναι τοποθετημένο στο εσωτερικό της κεντρικής μονάδας). Τα συστήματα Micropolis και Conner παρέχονται σε εξωτερικές μονάδες, ενώ διαθέσιμο και σε εξωτερική μονάδα είναι το σύστημα Compaq.

Το σύστημα Compaq συνδέεται στην κεντρική μονάδα μέσω ενός SCSI HBA δύο καναλιών, (Smart Array Controller). Ο HBA ελέγχεται από ένα μικροεπεξεργαστή Cyrix 486SLC και διαθέτει κρυφή μνήμη, (cache), η οποία τροφοδοτείται από μπαταρία. Ετσι, σε περίπτωση που διακοπεί η τροφοδοσία του συστήματος, τα δεδομένα που βρίσκονταν στην κρυφή μνήμη δεν χάνονται. Μπορεί να υποστηρίξει τα επίπεδα RAID 0, 1, 4 και 5. Οι μονάδες δίσκων που δεν χρησιμοποιούνται μπορούν να οριστούν στο σύστημα ως **on-line spares**.

Η δυνατότητα on-line spares αποτελεί μία βελτίωση της **hot-swap** η οποία παρέχονταν σε παλαιότερα συστήματα RAID. Οι hot-swappable συστοιχίες έδιναν την δυνατότητα αντικατάστασης ενός ελαττωματικού δίσκου κατά την διάρκεια της λειτουργίας τους. Στις συστοιχίες που προβλέπουν on-line spares, εφεδρικοί δίσκοι μπορούν εκ των προτέρων να τοποθετηθούν μέσα στο σύστημα. Σε περίπτωση δυστοκίας μίας συσκευής, τα on-line spares χρησιμοποιούνται αυτόματα από τον διαχειριστή της συστοιχίας ο οποίος γνωρίζει την ύπαρξη τους.

Το μοναδικό μειονέκτημα το οποίο εντοπίστηκε στην συστοιχία Compaq ήταν ο σημαντικά μεγάλος χρόνος ανακατασκευής δίσκου. Οταν το σύστημα βρισκόταν σε αδράνεια, ο χρόνος αυτός έφτασε την 1 ώρα για χωρητικότητα δίσκου της τάξης του 1GB. Αν, κατά την ανακατασκευή, το σύστημα δεν βρισκόταν σε αδράνεια, ο χρόνος απόκρισης του δεν επηρεαζόταν σημαντικά, ο χρόνος ανακατασκευής, όμως, για το ίδιο μέγεθος δίσκου, έφτασε τις 3 ώρες.

Στο σύστημα IBM 95, η συστοιχία περιέχεται στην βασική μονάδα και συνδέεται με αυτή μέσω ελεγκτή SCSI 2 καναλιών, (το οποίο χειρίζεται ένας RISC μικροεπεξεργαστής i960). Υποστηρίζεται η δυνατότητα hot-swapping. Το σύστημα συνοδεύεται από ειδικευμένο λογισμικό, για το NOVELL NetWare, μέσω του οποίου είναι δυνατή η παρακολούθηση της λειτουργίας του. Σε περίπτωση αλλαγής δίσκου, μέσω του παραπάνω λογισμικού, είναι δυνατή η εκκίνηση της ανακατασκευής των δεδομένων στην νέα συσκευή. Επίσης το παραπάνω λογισμικό παρέχει στοιχεία για τον φόρτο του συστήματος. Η ανακατασκευή ενός δίσκου χωρητικότητας 1GB διήρκεσε 20' ενώ, ταυτόχρονα, εξελισσόταν και άλλες λειτουργίες I/O.

Το σύστημα Micropolis RAIDION, [API95], αποτελεί εξωτερική μονάδα, (external unit), με ανεξάρτητη από την βασική μονάδα τροφοδοσία ισχύος. Η κάθε συσκευή δίσκου διαθέτει ανεξάρτητη από τις υπόλοιπες τροφοδοσία καθώς και ιδιαίτερο σύστημα εξαερισμού. Οι βλάβες που ενδεχομένως να υποστεί μία μεμονωμένη συσκευή δεν επηρεάζουν κατά κανένα απολύτως τρόπο τις υπόλοιπες. Σε παλαιότερες εκδόσεις της συστοιχίας, για την σύνδεση της με την κεντρική μονάδα χρησιμοποιείται ένας SCSI ελεγκτής της Adaptec.

Οταν εμφανιστεί πρόβλημα σε μία συσκευή δίσκου, το λογισμικό ελέγχου της συστοιχίας στέλνει μήνυμα στον διαχειριστή ενώ παρέχεται ανάλογη ένδειξη στην κονσόλα του συστήματος. Η συστοιχία ανακατασκευάζει, χωρίς να αποθηκεύει, τα κατεστραμμένα δεδομένα, όταν υπάρχουν αιτήσεις για αυτά, (on the fly reconstruction), μέχρις ότου αντικατασταθεί η προβληματική συσκευή. Η ανακατασκευή ενός δίσκου χωρητικότητας 1,5 GB διαρκεί περίπου 15'.

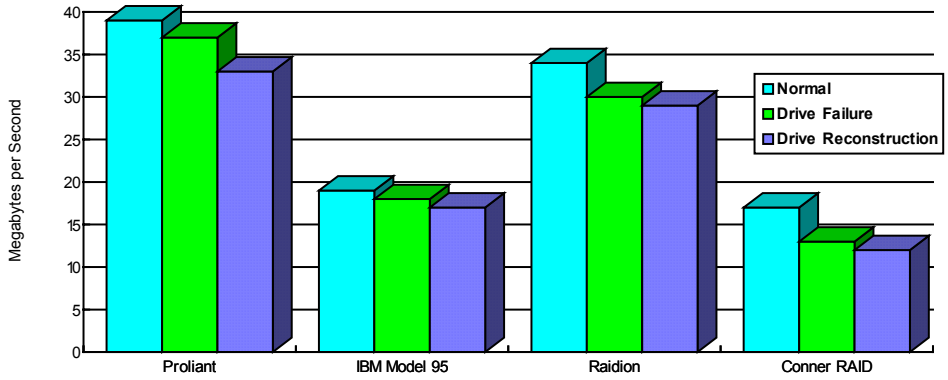
Η συστοιχία Raidion LTX είναι πλήρως scalable. Μπορεί να δεχτεί από 2 μέχρι 8 συσκευές δίσκου τις οποίες διαχειρίζεται ο Micropolis Gandiva RAID controller, (Fast & Wide SCSI II). Το μόνο πρόβλημα το οποίο εντοπίστηκε στην συστοιχία είναι η απουσία προστασίας στην τροφοδοσία του ελεγκτή, η οποία προκαλεί μεμονωμένο σημείο πιθανής βλάβης, (single point of failure), από το οποίο εξαρτάται πλήρως το όλο σύστημα. Υποστηρίζονται τα επίπεδα RAID 0, 1 και 5, η δυνατότητα hot-swapping καθώς επίσης και η δυνατότητα on-line spares. Επίσης μπορεί να ρυθμιστεί το stripping unit με το οποίο πρόκειται να λειτουργήσει η συστοιχία. Μέσω της ρύθμισης του stripping unit μπορεί να επιτευχθεί βελτιστοποίηση της απόδοσης της συσκευής για συγκεκριμένους τύπους εφαρμογών.

Για τον υπολογισμό των δεδομένων ισοτιμίας, (parity), χρησιμοποιείται ειδικός μικροεπεξεργαστής RISC 32 bits. Ο controller Gardiva διαθέτει 8 Mbytes μνήμης RAM την οποία χρησιμοποιεί για το λογισμικό ελέγχου της συστοιχίας, (≈1.5 MB), καθώς και για κρυφή μνήμη, (cache memory).

Το σύστημα Conner αποτελεί εξωτερική μονάδα RAID. Η συστοιχία συνοδεύεται από ένα SCSI controller της BusLogic, για την σύνδεση της με την κεντρική μονάδα, καθώς επίσης και από ειδικό λογισμικό για την εγκατάσταση, ρύθμιση και λειτουργία της. Μπορεί να υποστηρίξει τα επίπεδα 0, 1, 4 και 5. Οι επιδόσεις του συστήματος ήταν οι χειρότερες από όλα τα υπόλοιπα συστήματα τα οποία παρουσιάστηκαν παραπάνω και περιορίστηκαν ακόμη περισσότερο με την

ανακατασκευή δίσκου. Το σύστημα παρέχει την δυνατότητα on-line spares ενώ οι διαθέσιμες θέσεις δίσκων είναι 6.

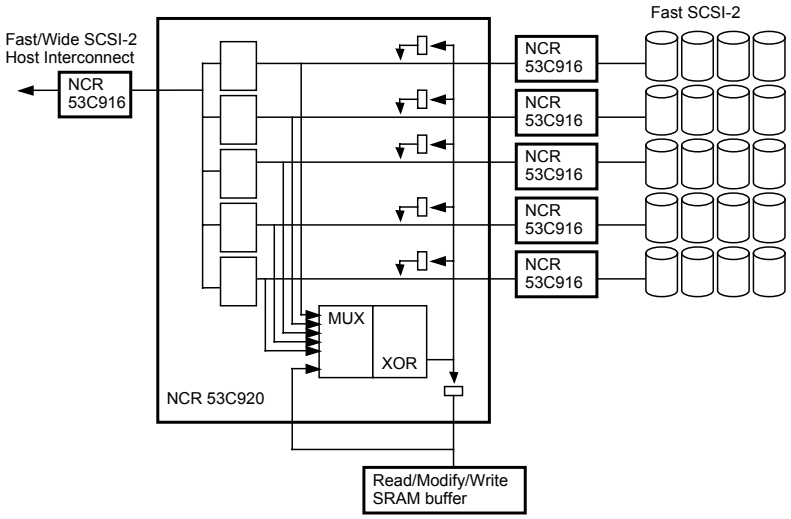
Οι επιδόσεις των παραπάνω συστημάτων, σε τρεις συγκεκριμένους τρόπους λειτουργίας, (Normal, Drive Failure και Drive Reconstruction), παρουσιάζονται στο διάγραμμα που ακολουθεί, (Σχήμα 13).



Σχήμα 13: Επιδόσεις συστημάτων RAID 5

7.2 NCR 6298

Το **NCR 6298 Disk Array Subsystem**, [CHE94], αποτελεί ένα σύστημα RAID χαμηλού σχετικά κόστους που υποστηρίζει τα επίπεδα 0, 1, 3 και 5 και είναι εμπορικά διαθέσιμο από το 1992. Το σύστημα έχει σχεδιαστεί για περιβάλλοντα εμπορικών εφαρμογών. Υποστηρίζει μέχρι και 4 ελεγκτές, πλεονάζοντα στοιχεία τροφοδοσίας και εξαερισμού καθώς και 20 μονάδες σκληρού δίσκου, SCSI II σε διαστάσεις 3.5". Ολα τα συστατικά του συστήματος μπορούν να αντικατασταθούν ενώ αυτό βρίσκεται σε λειτουργία και εξυπηρετεί αιτήσεις. Δεν υπάρχει η δυνατότητα on-line spares. Στο Σχήμα 14 που ακολουθεί δίνεται ένα block διάγραμμα της συστοιχίας.



Σχήμα 14: Block διάγραμμα NCR 6298

Ο ελεγκτής της συστοιχίας χαρακτηρίζεται από μία βηματική αρχιτεκτονική, (**lock-step**), η οποία εξαλείφει την ανάγκη buffering. Για όλες τις αιτήσεις πλην των εγγραφών επιπέδου 5, τα δεδομένα οδηγούνται μέσω του ελεγκτή απευθείας στις συσκευές. Ο ελεγκτής είναι υπεύθυνος για τον διπλασιασμό των δεδομένων σε περίπτωση mirroring καθώς και για τον υπολογισμό της πληροφορίας ισοτιμίας σε περίπτωση επιπέδου 3. Οι παραπάνω λειτουργίες πραγματοποιούνται ταυτόχρονα με την μεταφορά των δεδομένων.

Στο επίπεδο 5 δεν υποστηρίζονται εγγραφές που να καλύπτουν ένα πλήρες stripe unit. Χρησιμοποιείται ένας ενδιάμεσος buffer στατικής RAM. Οταν υπάρξει αίτηση εγγραφής, τα παλιά δεδομένα και η ισοτιμία διαβάζονται από το δίσκο, (lock-step), υπολογίζεται το XOR τους και αποθηκεύονται σε ένα parity buffer χωρητικότητας 64 KB. Τα δεδομένα που προέρχονται από την κεντρική μονάδα, (host), γίνονται XOR με το περιεχόμενο του buffer για τον υπολογισμό της νέας ισοτιμίας, (up-to-date parity). Θα πρέπει να τονιστεί ότι η αρχιτεκτονική δεν επιτρέπει την χρονική επικάλυψη, (overlap), στην μεταφορά δεδομένων.

Ο βηματικός τρόπος λειτουργίας χρησιμοποιείται και στην ανακατασκευή δίσκου. Τα δεδομένα διαβάζονται σύγχρονα από τις συσκευές που παραμένουν λειτουργικές, υπολογίζεται το XOR τους και ξαναγράφονται στον δίσκο που αντικατέστησε τον προβληματικό. Η ανακατασκευή στο σύστημα είναι ιδιαίτερα γρήγορη και προσεγγίζει τον χρόνο εγγραφής ενός δίσκου.

Η σύνδεση με την βασική μονάδα, (host), ακολουθεί το πρότυπο Fast & Wide SCSI-2 με ρυθμούς της τάξης των 20MB/s. Για τα κανάλια των δίσκων ακολουθείται το πρότυπο Fast, Narrow SCSI-2 με ρυθμό 10MB/s. Εξαιτίας της βηματικής

αρχιτεκτονικής ο ρυθμός προς την κεντρική μονάδα περιορίζεται στα 10MB/s όταν το σύστημα λειτουργεί σε επίπεδο 0, 1 ή 5.

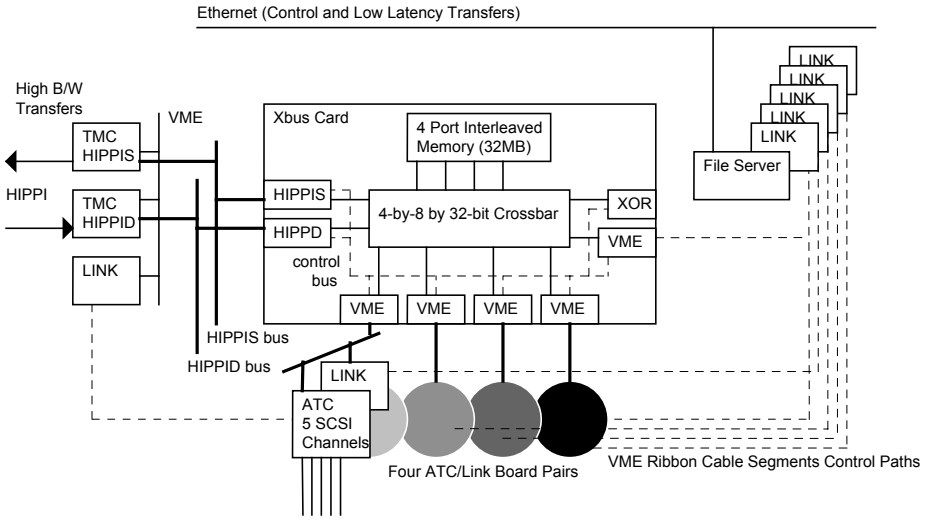
Η χωρητικότητα της συστοιχίας μπορεί να αυξηθεί στα 21Gb ενώ τα λειτουργικά συστήματα που συνεργάζονται μαζί της είναι τα UNIX, Windows NT και LAN Manager.

7.3 RAID-II STORAGE SERVER

Το σύστημα RAID II, [KAT93], [CHE94] αποτελεί ένα εξυπηρετητή αρχείων δικτύου με μεγάλο αποδιδόμενο εύρος ζώνης. Εχει σχεδιαστεί και υλοποιηθεί στο Πανεπιστήμιο Berkeley, California σε μία προσπάθεια να διερευνηθούν τα συστήματα περιφερειακής αποθήκευσης που χαρακτηρίζονται από υψηλές χωρητικότητες, αξιοπιστία και επιδόσεις.

Το σύστημα συνδέει μία συστοιχία δίσκων SCSI σε ένα δίκτυο **HIPPI**, (High Performance Parallel Interface). Ενα από τα πλέον σημαντικά χαρακτηριστικά του RAID-II είναι η δυνατότητα του να παρέχει υπηρεσίες δικτύου στο δίκτυο που συνδέθηκε χωρίς να απαιτείται εμπλοκή, στην μεταφορά των δεδομένων, του συμβατικού εξυπηρετητή αρχείων, (Sun4/280 workstation), ο οποίος εμφανίζει σημαντικά χαμηλότερες επιδόσεις. Για να επιτευχθεί η παραπάνω δυνατότητα, σχεδιάστηκε και υλοποιήθηκε μία ειδική πλακέτα τυποποιημένου κυκλώματος που ονομάζεται κάρτα Xbus.

Η κάρτα Xbus παρέχει ένα μονοπάτι δεδομένων, (data path), μεγάλου εύρους ζώνης το οποίο συνδέει τα βασικότερα συστατικά του συστήματος: το δίκτυο HIPPI, 4 διαύλους VME και μία interleaved, multiported μνήμη. Η κάρτα Xbus ενσωματώνει μία ειδική διάταξη για τον υπολογισμό της πληροφορίας ισοτιμίας, (parity computation engine). Η σύνδεση μεταξύ των συστατικών επιτυγχάνεται μέσω ενός μεταγωγέα crossbar 4×8, ο οποίος μπορεί να ανταπεξέλθει σε ρυθμούς της τάξης των 160MB/s. Το όλο σύστημα ελέγχεται από ένα εξωτερικό εξυπηρετητή αρχείων Sun 4/280 μέσω κατάλληλου προσαρμοστικού, (memory mapped control interface). Στο Σχήμα 15 που ακολουθεί παρέχεται ένα block διάγραμμα του ελεγκτή RAID-II.



Σχήμα 15: Αρχιτεκτονική RAID-II

Ενας συγκεντρωτικός πίνακας για εμπορικά διαθέσιμα προϊόντα RAID παρατίθεται στην συνέχεια, (Πίνακας 6):

Πίνακας 6: Εμπορικά διαθέσιμα προϊόντα RAID

Κατασκευαστής	Μοντέλο	Λειτουργικά Συστήματα	RAID επίπεδα
Compaq		Netware 3.x & 4.0	4, 5
Conner Peripherals	CR611E/CR611DM	Netware	0, 1, 5
Core	LAN Array LA/MA	Netware, Lan Manager, OS/2, Vines	3, 5
Data General	CLARiiON series 2000	All	0, 1, 3, 5
DEC	StorageWorks RAID	Solaris, SCO Unix, Netware	0, 1, 5
Hewlett-Packard		Netware, Unix	5
Micropolis	RAIDION LS/LT	Netware 3.11	5
Mylex	DAC960	Unix, Netware	0, 1, 5
NCR	NCR 6298	Unix, Windows NT, Lan Manager	0, 1, 3, 5
Quantum		Netware 3.x	1, 3, 5
StorageDimensions		Netware 3.x & 4.0	5

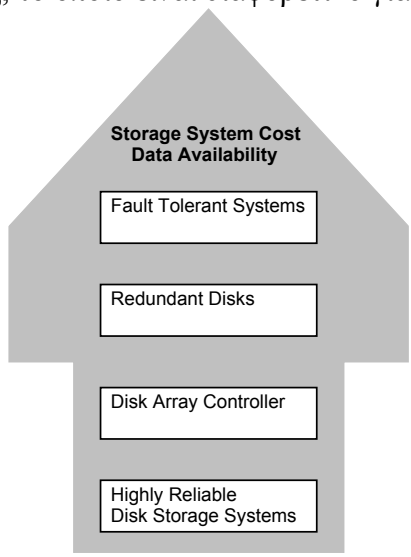
Οπως φαίνεται από τον παραπάνω πίνακα, υπάρχει μία πληθώρα προϊόντων RAID που μπορούν να λειτουργήσουν σε πολλά διαφορετικά επίπεδα. Πολλά από τα συστήματα αυτά καλύπτουν τόσο το επίπεδο 1 (disk mirroring) όσο και το 0 (disk stripping). Ολοι όμως οι κατασκευαστές κινούνται προς το επίπεδο 5 το οποίο φαίνεται να είναι και το επικρατέστερο.

8. ΣΧΕΣΗ ΚΟΣΤΟΥΣ ΕΞΟΠΛΙΣΜΟΥ - ΔΙΑΘΕΣΙΜΟΤΗΤΑΣ ΔΕΔΟΜΕΝΩΝ

Στο κεφάλαιο αυτό συγκρίνεται η διαθεσιμότητα των δεδομένων που αποθηκεύονται σε συστοιχίες δίσκων με άλλες λύσεις που προσφέρει η σύγχρονη τεχνολογία πληροφορικής. Ενας καλός τρόπος για να αντιμετωπιστεί η έννοια της διαθεσιμότητας στα συστήματα αυτά αφορά στην θεώρηση μίας ιεραρχίας για την προστασία των δεδομένων. Σε αυτό το παράδειγμα, υψηλότερη διαθεσιμότητα επιτυγχάνεται μέσω τεχνικών που οδηγούν σε σημαντικά υψηλότερα κόστη παραγωγής, (Γενικός Κανόνας Συστημάτων Fault Tolerance, [CRI91]). Για κάθε επιθυμητό επίπεδο διαθεσιμότητας θα πρέπει να επιλεγεί μία συγκεκριμένη λύση αποθήκευσης, (storage).

Στο Σχήμα 16 παρουσιάζονται πολλά επίπεδα διαθεσιμότητας, ξεκινώντας από συστήματα αποθήκευσης με δίσκους υψηλής αξιοπιστίας, (high reliability), και

πηγαίνοντας προς τα πάνω σε όλο και αυξανόμενα επίπεδα διαθεσιμότητας δεδομένων. Οπως είναι αναμενόμενο το κόστος αυξάνει κατά την μετακίνηση προς τα πάνω τμήματα της πυραμίδας του σχήματος 16, [PAV92]. Σε ορισμένες εφαρμογές, η απαίτηση για υψηλή αξιοπιστία είναι καταλυτική, οπότε στην σχετική αγορά εξοπλισμού δεν λαμβάνεται διόλου υπόψη το κόστος. Ο πλέον σημαντικός παράγοντας στην επιλογή λύσεων υψηλής διαθεσιμότητας είναι το κόστος του χρόνου downtime του συστήματος, το οποίο είναι διαφορετικό για κάθε περίπτωση.



Σχήμα 16: Ταξινόμηση συστημάτων περιφερειακής αποθήκευσης ανάλογα με το κόστος και την διαθεσιμότητα που παρέχουν

Κοντά στην κορυφή της ιεραρχίας υπάρχουν τα συστήματα δίσκων που χρησιμοποιούν πλήρη πλεονασμό, (Redundant Disks). Κάποια από αυτά διπλασιάζουν πλήρως την αποθήκευση, (κατοπτρικοί δίσκοι), ενώ ταυτόχρονα χρησιμοποιούν διπλούς ανεμιστήρες, παροχές ισχύος, ελεγκτές, καλώδια και HBA. Με αυτόν τον τρόπο εξαλείψουν τον κίνδυνο μεμονωμένου σημείου πιθανής βλάβης, (single point of failure). Επιπρόσθετα χαρακτηριστικά ανοχής δυσλειτουργιών, (fault tolerance), μπορούν να προστεθούν στα συστήματα αυτά για την επίτευξη ακόμη μεγαλύτερων βελτιώσεων στην αξιοπιστία και την διαθεσιμότητα, (μηδενικό downtime). Σε άλλες εφαρμογές παρατηρείται ακόμη και ο τριπλασιασμός των κρισίμων συστατικών, (triple redundancy).

Στο ενδιάμεσο επίπεδο της ιεραρχίας τοποθετούνται οι συστοιχίες δίσκων που παρέχουν προστασία στα δεδομένα. Αυτό το επίπεδο χαρακτηρίζεται από υψηλή διαθεσιμότητα του συστήματος με χαμηλό κόστος σε σχέση με αυτό του πλήρη πλεονασμού. Οι συστοιχίες δίσκων δεν διπλασιάζουν το κάθε συστατικό τους, αλλά ο κύριος στόχος τους είναι να παρέχουν προστασία από τις βλάβες δίσκων που ενδεχομένως συμβούν.

Σε αυτήν την κατηγορία ανήκουν και τα επίπεδα RAID 0-6, τα οποία πλεονεκτούν από πλευράς διαθεσιμότητας/κόστους από τα συστήματα του πλήρη πλεονασμού.

ΒΙΒΛΙΟΓΡΑΦΙΑ - ΑΝΑΦΟΡΕΣ

Alford R. 1992, Disk Arrays Explained, BYTE (October), Vol. 17, No. 10

Apiki S. 1995, Simple Scalable RAID, BYTE (February), Vol. 20, No. 2

Bitton D. and Gray J. 1988, Disk Shadowing, Proceedings of the 14th VLDB Conference Los Angeles, California.

Chen P., Lee E., Gibson G., Katz R. and Patterson D. 1994, RAID: High-Performance, Reliability Secondary Storage, ACM Computing Surveys (June), Vol.26, No. 2

Chen P. and Lee E. 1993, Striping in a RAID Level 5 Disk Array, Tech.Report CSE-TR-181-93, University of Michigan, Ann Arbor. Mich.

Cristian F. 1991, Understanding Fault-Tolerant Distributed Systems, Communications of the ACM (February), Vol. 34, No.2

Katz P., Chen P., Drapeau A., Lee E., Lutz K., Miller E., Seshan S. and Patterson D. 1993, RAID-II: Design and Implementation of a Large Scale Disk Array Controller, In the 1993 Symposium on Integrated Systems. MIT Press, Cambridge, Mass.

LAN MAGAZINE 1994, RAM RAIDers, (July).

Lawrence B. 1992, No More Data Loss, BYTE (August), Vol. 17, No. 8

Lee E. and Katz R. 1991, Performance Consequences of Parity Placement in Disk Arrays, Proceedings of the 4th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS IV). IEEE, New York

Patterson D., Gibson G. and Katz R. 1988, A case for redundant arrays of inexpensive disks (RAID). In International Conference on Management of Data (SIGMOD). ACM New York

Pavlinik Ed 1992, Disk Arrays and RAID, Datapro (September), McGraw Hill

Tanenbaum A. 1992, Modern Operating Systems, Prentice-Hall

Wallace Scott 1994 Managing Mass Storage, BYTE (March), Vol. 19, No. 3

Μπάλιος Ζ., Παλάζης Β. και Σκούρα Α. 1992, Αξιοπιστία Υπολογιστικών Συστημάτων - Μελέτη και μοντέλα αξιοπιστίας λογισμικού, Πτυχιακή Εργασία, Τμήμα Πληροφορικής Πανεπιστημίου Αθηνών.

Μανωλόπουλος 1991, Δομές Δεδομένων Τόμος Β', Art of Text

Σωτηρόπουλος Τ. 1994, RAID Πλενάζουσες Συστοιχίες Ανέξοδων Δίσκων, OPEN NEWSLETTER (Μάϊος), Ετος 2ο, Φύλλο 5ο