ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΗΠΕΙΡΟΥ
EPIRUS INSTITUTE OF TECHNOLOGY
ARTA , ZIP CODE : 47100 , TEL NUMBER : +30 26810 50000

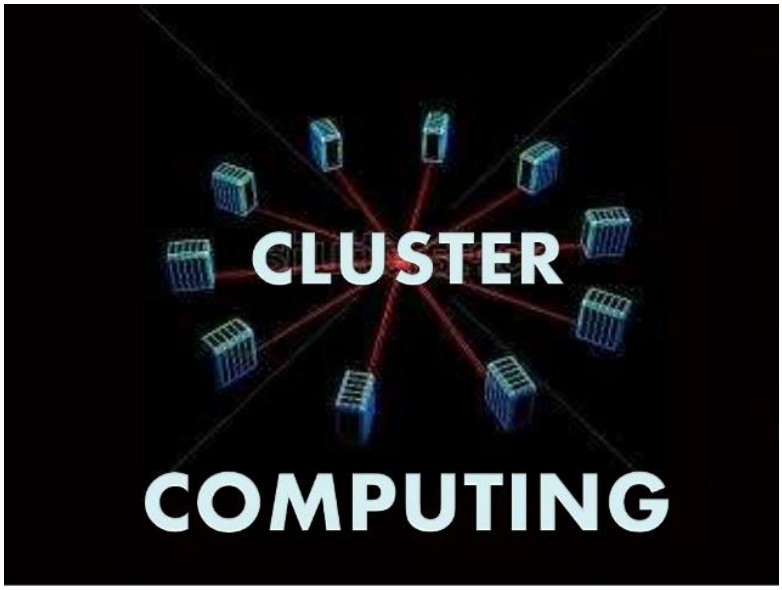
ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΗΠΕΙΡΟΥ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΥΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Δημιουργία Cluster σε εικονικό περιβάλλον»



Παναγιώτου Αλέξανδρος

Α.Μ: 8299

Επιβλέπων Καθηγητής: Λιαροκάπης Δημήτριος

Αρτα, Ιούνιος 2016

Δήλωση Πνευματικής Ιδιοκτησίας

Η παρούσα Πτυχιακή εργασία αποτελεί προϊόν αποκλειστικά δικής μου προσπάθειας. Όλες οι πηγές που χρησιμοποιήθηκαν περιλαμβάνονται στη βιβλιογραφία και γίνεται ρητή αναφορά σε αυτές μέσα στο κείμενο όπου έχουν χρησιμοποιηθεί.

ΥΠΟΓΡΑΦΕΣ

ΕΥΧΑΡΙΣΤΙΕΣ

Ευχαριστώ θερμά τον επιβλέπον καθηγητή μου, κ. Λιαροκάπη για την εποικοδομητική συνεργασία και καθοδήγηση κατά τη διάρκεια εκπόνησης της εργασίας. Επίσης θα ήθελα να ευχαριστήσω θερμά τους συμφοιτητές και φίλους μου από το Πανεπιστήμιο της Αλμερίας UAL, (Universidad de Almeria) του τμήματος πληροφορικής, Andres Rafael Rojas Garcia, Alfonso Tejedor Moreno, Sergio Martin και Vicror Suarez Garcia, καθώς και από το τμήμα Ψυχολογίας του Πανεπιστημίου της Αλμερίας την Purificacion de Lourdes Ceballos Jimenez, για την πολύτιμη προσφορά τους σε πηγές και σε υλικό , καθώς και στην ορθότερη μετάφραση πληροφοριών από τα Ισπανικά.

Περιεχόμενα

Κεφάλαιο 1

Κεφάλαιο 2

Περίληψη

Σε ένα κόσμο όπου η πληροφορία αποτελεί πλεονέκτημα και ο όγκος της αυξάνεται διαρκώς χάρη στη ραγδαία ανάπτυξη της τεχνολογίας, αναζητούνται νέοι δρόμοι για την επίτευξη υψηλότερων επιδόσεων, με χαμηλό κόστος χρήσης και λειτουργίας, αλλά και μεγαλύτερη ασφάλεια των δεδομένων. Σκοπός μας είναι να δείξουμε έναν από αυτούς τους «δρόμους» που ακούει στο όνομα «συστοιχία υπολογιστών» (Computer Cluster) και είναι ένας από τους ποιο διαδεδομένους τρόπους λειτουργείας των λεγόμενων «υπερυπολογιστών» (Supercomputers). Αν και αυτή η γνώση υπάρχει εδώ και αρκετά χρόνια, δεν χρησιμοποιείτε πέραν από κάποια ειδικά ερευνητικά κέντρα για επεξεργασία δεδομένων της Αστρονομίας, Μετεωρολογίας, Γεωλογίας και σε Στρατιωτικές έρευνες, τέλος υπάρχουν και ορισμένες Πανεπιστημιακές κοινότητες καθώς και ερασιτέχνες που πειραματίζονται με την συγκεκριμένη τεχνολογία.

Η συγκεκριμένη τεχνολογία θα μπορούσε να βοηθήσει μικρές επιχειρήσεις αλλά και μεγάλες, καθώς και Πανεπιστημιακά ιδρύματα να γλυτώσουν ένα σημαντικό κεφάλαιο, από την αγορά και την ανανέωση των υπολογιστικών τους συστημάτων, απλά αξιοποιώντας το είδη υπάρχον υλικό με αποδοτικότερο τρόπο, βελτιστοποιώντας και τις επιδώσεις των συστημάτων τους αλλά και την ασφάλεια των δεδομένων τους και όλα αυτά με ένα εξαιρετικά χαμηλό κόστος και φιλικότερο τρόπο προς περιβάλλον, καθώς η συγκεκριμένη τεχνολογία αποτελεί ένα τρόπο «ανακύκλωσης» των πόρων μας και επαναχρησιμοποίησης.

Γι’ αυτό λόγο κατασκευάσαμε ένα cluster σε εικονικό περιβάλλον με την χρήση του VirtualBox της Oracle και πήραμε τα σχετικά γραφήματα απόδοσης, με την βοήθεια του Ganglia και του SuperPi. Τα παραπάνω περιγράφονται στις επόμενες σελίδες αναλυτικά για να μπορέσουμε να κατανοήσουμε τι είναι το cluster, που χρησιμοποιείτε, τις τεχνικές του προδιαγραφές, τα Λειτουργικά Συστήματα που υποστηρίζουν την συγκεκριμένη τεχνολογία καθώς και τον τρόπο κατασκευής και χρήσεις.

ΚΕΦΑΛΑΙΟ 1

High Performance Computing (HPC)

1.1 Εισαγωγή

Με τον όρο High Performance Computing (HPC) εννοούμε όλους τους υπερυπολογιστές (Super computers) και τεχνικές παράλληλης επεξεργασίας για να μπορούν να λύσουν πολυσύνθετα προβλήματα. Ο όρος Υπολογιστές Υψηλών Επιδόσεων (High Performance Computing) και ο όρος υπερυπολογιστές χρησιμοποιούνται πολλές φορές εναλλάξ. Η HPC τεχνολογία επικεντρώνονται στην ανάπτυξη παράλληλων αλγορίθμων επεξεργασίας και συστημάτων, τόσο με παράλληλες υπολογιστικές τεχνικές όσο και με διαχειριστικές.

Οι Υπολογιστές Υψηλών Επιδόσεων (HPC), χρησιμοποιούνται συνήθως για πολύ απαιτητικές αναλύσεις και προσημειώσεις π.χ. της συμπεριφοράς των αστεριών ενός γαλαξία ή της ατμόσφαιρας σε πλανητική κλίμακα. Τα HPC συστήματα έχουν τη δυνατότητα να εξασφαλίσουν σταθερή απόδοση μέσα από την ταυτόχρονη χρήση των υπολογιστικών πόρων.

Οι υπερυποογιστές εισήχθησαν τη δεκαετία του 1960 με πρώτο τον υπερυπολογιστή Atlas στο Πανεπιστήμιο του Manchester από τον Seymour Cray (Αμερικανός ηλεκτρολόγος μηχανικός) από την εταιρία κατασκευής υπολογιστών Control Data Corporation (CDC).Το 1976 ο Cray-1 των 80Mhz έγινε ο ποιο επιτυχημένος υπερυπολογιστής στην Ιστορία.[1][2] Οι υπερυπολογιστές της δεκαετίας του 1970 χρησιμοποιούσαν λίγους μόνο επεξεργαστές σε αντίθεση με την δεκαετία του 1990 όπου άρχισαν να εμφανίζονται μηχανήματα με χιλιάδες επεξεργαστές.[3][4] Χαρακτηριστικό παράδειγμα ο υπερυπολογιστής της Fujitsu Numerical Wind Tunel το 1994 όπου έφερε 166 επεξεργαστές και έφτανε την ταχύτητα των 1.7 Gigaflops ανα επεξεργαστή καθώς στα τέλη του 20^{ου} αιώνα εμφανίστηκαν συστοιχίες παράλληλων υπερυπολογιστών με δεκάδες χιλιάδες επεξεργαστές.[5] Η ικανότητα υπολογισμών μετριέται συνήθως με τον όρο Flops (Floating-

Εικόνα: 1.1 Υπερυπολογιστής Cray-1

point Operations Per Second, υπολογισμοί κινητής υποδιαστολής ανά δευτερόλεπτο). Η υπολογιστική ικανότητα των σημερινών υπερυπολογιστών έχει ξεπεράσει το 1 PetaFlop.[6][7]

Ο ποιο δυνατός υπερυπολογιστής μέχρι και το 2014 είναι ο Tianhe-2 (Milky Way 2)[8][9] Ο συγκεκριμένος υπερυπολογιστής βρίσκεται και χρησιμοποιείται Από το εθνικό κέντρο υπερυποογιστών στο Guangzhou στην Κίνα. Ο υπερυπολογιτής διαθέτει 3.120.000 πυρήνες με απόδοση 33.86 petaflops. Η μηχανή αναπτύχθηκε από το Εθνικό Πανεπιστήμιο Αμυντικής Τεχνολογίας της Κίνας (NUDT). Εκτός από την αμυντική χρήση του, χρησιμοποιείτε και για πρόβλεψη σεισμών καθώς και για την παρακολούθηση των κλιματικών αλλαγών.[10][11][12]

1.2 Χρήση ευρυζωνικών δικτύων

Αν και με την πάροδο του χρόνου το κόστος των υπερυπολογιστών μειώθηκε σημαντικά από την άλλη οι ανάγκες για επεξεργασία ολοένα και περισσότερων αλλά και μεγαλύτερων σε όγκο δεδομένων αυξάνονταν, ιδίως με την χρήση πληροφοριακών συστημάτων μηχανογράφησης σε εταιρίες και την είσοδο στο internet. Έτσι αναζητήθηκαν από εταιρίες μεσαίου μεγέθους αλλά και από Πανεπιστημιακά ιδρύματα τα οποία χρειάζονταν αυξημένη υπολογιστική ισχύ αλλά παράλληλα και φθηνή, διαφορετική τρόποι αξιοποίησης των μέσων όπου διαθέτουν. [13]

Ένας τρόπος ήταν με την εξάπλωση των ευρυζωνικών δικτύων πολλοί οικιακοί υπολογιστές μπορούν να συνδεθούν σε διακομιστές μέσω Διαδικτύου και λειτουργώντας αθροιστικά να επιτύχουν υψηλή υπολογιστική ισχύ (υπολογιστικό πλέγμα) αντίστοιχη των υπερυπολογιστών . Δύο σημαντικές προσπάθειες προς την κατεύθυνση αυτή είναι η πλατφόρμα ενδιάμεσου λογισμικού BOINC του πανεπιστημίου Berkeley της Καλιφόρνιας και το Folding@Home του Πανεπιστημίου Stanford,[14] με χρήση των επεξεργαστών Cell του Playstation 3. Το BOINC είναι πολυδιάστατο εγχείρημα με βασική εφαρμογή το SETI@Home. Η συνολική μέση ταχύτητά του είναι 550 TeraFlops.[15][16]

1.2.1 Grid computing

Grid computing είναι η συλλογή των πόρων των υπολογιστών από πολλαπλές τοποθεσίες για την επίτευξη ενός κοινού στόχου. Το πλέγμα (Grid) μπορεί να θεωρηθεί ως ένα κατανεμημένο σύστημα με μη-διαδραστικό φόρτο εργασίας που αφορά ένα μεγάλο αριθμό αρχείων.[17] Το Grid computing διακρίνεται από τα συμβατικά συστήματα High Performance Computing, όπως το cluster computing στο ότι τα δίκτυα τείνουν να είναι πιο ετερογενείς και

γεωγραφικά διεσπαρμένα (έτσι δεν είναι φυσικά σε ένα μέρος). Παρά το γεγονός ότι ένα ενιαίο πλέγμα μπορεί να διατεθεί για μια συγκεκριμένη εφαρμογή, συνήθως ένα πλέγμα χρησιμοποιείται για διάφορους σκοπούς. Τα πλέγματα κατασκευάζονται συχνά με γενικής χρήσης βιβλιοθήκες λογισμικού grid middleware .

Το μέγεθος του grid ποικίλλει. Το grid είναι μια μορφή κατανεμημένων υπολογστων, στην οποία ένας "υπερ-εικονικός υπολογιστής» αποτελείται από πολλούς δικτυωμένους υπολογιστές σε χαλαρή σύνδεση που ενεργούν από κοινού για την εκτέλεση μεγάλων εργασιών. Για ορισμένες εφαρμογές, "distributed" ή "grid" computing, μπορεί να θεωρηθεί ως ένα ιδιαίτερο είδος των παράλληλων υπολογιστών που βασίζεται σε πλήρεις υπολογιστές (με επεξεργαστές, μνήμη, τον αποθηκευτικό χώρο , τροφοδοτικά, κάρτες δικτύου κτλ.) που συνδέονται σε ένα δίκτυο (ιδιωτικό , δημόσιο ή το Διαδίκτυο) με μια συμβατική διασύνδεση δικτύου, όπως Ethernet. Αυτό έρχεται σε αντίθεση με την παραδοσιακή έννοια του υπερυπολογιστή, η οποία έχει πολλούς επεξεργαστές που συνδέονται με ένα τοπικό δίαυλο υψηλής ταχύτητας.[18]

1.2.2 Σύγκριση Grid computing με Supercomputers

"Κατανεμημένα συστήματα» ή «πλέγμα» υπολογιστών σε γενικές γραμμές είναι ένας ειδικός τύπος των παράλληλων συστημάτων υπολογιστών που βασίζεται σε πλήρεις υπολογιστές (με επεξεργαστές, μνήμη RAM, αποθηκευτικό χώρο, κάρτες γραφικών, τροφοδοτικά, κάρτες δικτύου κτλ.) που συνδέονται σε ένα δίκτυο (ιδιωτικό, δημόσιο ή το Διαδίκτυο) με μία συμβατική διασύνδεση δικτύου, όπου συνδυάζουν τους πόρους τους για την επίτευξη ενός σκοπού.

Το κύριο μειονέκτημα στις επιδόσεις των κατανεμημένων συστημάτων είναι ότι οι διάφορες περιοχές επεξεργασίας και τοπικής αποθήκευσης δεν έχουν συνδέσεις υψηλής ταχύτητας. Αυτή η διάταξη είναι έτσι καλά προσαρμοσμένη σε εφαρμογές στις οποίες πολλαπλοί παράλληλοι υπολογισμοί μπορεί να πραγματοποιηθούν ανεξάρτητα, χωρίς την ανάγκη να επικοινωνούν τα ενδιάμεσα αποτελέσματα μεταξύ των επεξεργαστών. Η high-end γεωγραφική επεκτασιμότητα διασκορπισμένων grid είναι γενικά ευνοϊκή, λόγω της χαμηλής ανάγκης για τη σύνδεση μεταξύ των κόμβων σε σχέση με την ικανότητα του δημόσιου Διαδικτύου.

Υπάρχουν επίσης ορισμένες διαφορές όσον αφορά τον προγραμματισμό και την ανάπτυξη. Μπορεί να είναι δαπανηρό και δύσκολο να γραφτούν προγράμματα που μπορεί να τρέξουν σε περιβάλλον ενός υπερυπολογιστή, που μπορεί να έχει ένα προσαρμοσμένο

λειτουργικό σύστημα, ή να απαιτούν το πρόγραμμα για την αντιμετώπιση θεμάτων συγχρονισμού. Αν ένα πρόβλημα μπορεί να παραλληλίζεται επαρκώς, από "thin" επίπεδο σε "grid" μπορεί να επιτρέψει σε συμβατικά, αυτόνομα προγράμματα να διανέμουν ένα διαφορετικό μέρος του ίδιου προβλήματος για να τρέξει σε πολλαπλά μηχανήματα. Αυτό καθιστά δυνατό να γραφτούν και να εντοπιστούν σφάλματα σε μια ενιαία συμβατική μηχανή, και εξαλείψει τις επιπλοκές που οφείλονται στις πολλαπλές εμφανίσεις του ίδιου προγράμματος που τρέχει στην ίδια κοινόχρηστη μνήμη και αποθηκευτικό χώρο την ίδια στιγμή.[19]

1.2.3 Επεξεργαστής Σάρωσης (CPU Scavenging)

CPU-scavenging ή cycle-scavenging δημιουργεί ένα «πλέγμα» από τους αχρησιμοποίητους πόρους σε ένα δίκτυο των συμμετεχόντων (είτε σε παγκόσμιο επίπεδο είτε στο εσωτερικό ενός οργανισμού). Συνήθως η τεχνική αυτή χρησιμοποιεί σε desktop υπολογιστές όπου στις παρακάτω περιπτώσεις οι αξιοποίηση των πόρων τους θα χανόταν όπως το βράδυ, κατά τη διάρκεια του γεύματος, ή ακόμα και στα διάσπαρτα δευτερόλεπτα κατά τη διάρκεια της ημέρας, όταν ο υπολογιστής είναι σε αναμονή για την είσοδο του χρήστη. Στην πράξη, οι συμμετέχοντες υπολογιστές δωρίζουν επίσης κάποια υποστήριξη όπως κάποιο ποσό του χώρου αποθήκευσης στο δίσκο, RAM, και το εύρος ζώνης του δικτύου, εκτός από την δύναμη της CPU όπου προσφέρει ούτος ή αλλιώς.

1.2.4 Εφαρμογή Grid Computing

Το SETI@home ("SETI at home") είναι μια διαδικτυακή βάση «εθελοντών» υπολογιστών που χρησιμοποιούν την πλατφόρμα λογισμικού BOINC, που φιλοξενείται από το Εργαστήριο Θετικών Επιστημών, στο Πανεπιστήμιο της Καλιφόρνιας, Berkeley, στις Ηνωμένες Πολιτείες. Το SETI είναι ένα ακρωνύμιο του Extra Terrestial Intelligence για την αναζήτηση εξωγήινης νοημοσύνης.

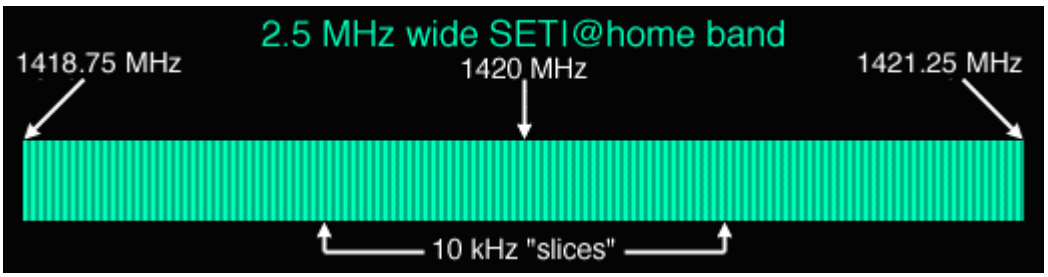
Εικόνα:1.2 Ραδιοτηλεσκόπιο Arecibo Puerto Rico

Σκοπός του είναι να αναλύσει ραδιοσήματα, ψάχνοντας για σημάδια εξωγήινης νοημοσύνης, και είναι μία από τις πολλές δραστηριότητες που αναλαμβάνονται στο πλαίσιο

του SETI. Για πρώτη φορά κυκλοφόρησε στο κοινό στις 17 Μαΐου του 1999 καθιστώντας το, το δεύτερο μεγάλης κλίμακας κατανεμημένο σύστημα πληροφορικής μέσω διαδικτύου. Αυτή η αναζήτηση βασίζεται στο να συλλέγουν-«ακούνε» μεταδόσεις ραδιοσημάτων από το διάστημα μέσο του ραδιοτηλεσκόπιου Arecibo στο Puerto Rico. [20]

Τα δεδομένα συλλέγονται ψηφιοποιούνται, αποθηκεύονται και αποστέλλονται στις εγκαταστάσεις του SETI@Home στο Berkeley της Καλιφόρνιας. Τα δεδομένα στην συνέχεια αναλύονται σε μικρά κομμάτια συχνοτήτων και χρόνου. Έπειτα θα πρέπει να εξετάσουν κάθε συχνότητα ξεχωριστά και να εξετάσουν αν μεταφέρει κάποιο «έξυπνο» μήνυμα ή απλά μεταφέρουν θόρυβο. Το πρόβλημα που δημιουργείτε είναι ότι ο όγκος των δεδομένων που καλούνται να επεξεργαστούν είναι τεράστιος. Για να πάρουμε μια ιδέα για την έκταση του προβλήματος αρκεί να δούμε πως η κεραία που χρησιμοποιεί το SETI@Home, εγγράφονται 35Gigabytes δεδομένων κάθε μέρα, τα οποία χρειάζονται όπως αναφέραμε και ποιο πάνω ιδική επεξεργασία βάση περίπλοκων αλγορίθμων. Σύμφωνα με την ιστοσελίδα του SETI@Home, ένας μέσος οικιακός υπολογιστής χρειάζεται από 10 ώρες έως και 50 ώρες για να τα επεξεργαστεί «μια μονάδα επεξεργασίας». Τα 35Gigabyte χωρίζονται σε 140.000 «μονάδες εργασίας». Αυτό αντιστοιχεί σε 4.200.000 ώρες επεξεργασίας δεδομένων μίας μόνο ημέρας![21]

Το κόστος αγοράς και συντήρησης ενός υπερυπολογιστή (HPC) που να κάνει αυτή την «δουλειά» θα ήταν απαγορευτικό για ένα πανεπιστήμιο. Γι' αυτό σκέφτηκαν έναν ποιο έξυπνο και οικονομικό τρόπο, να «σπάσουν» τα δεδομένα σε μικρότερα κομμάτια και μέσω της BOINC εφαρμογής Client-Server να αποστέλλουν τα πακέτα προς επεξεργασία μέσο internet σε εκατομμύρια οικιακούς υπολογιστές εθελοντές. Οι χρήστες όταν είναι συνδεδεμένοι



Εικόνα : 1.3 Επειδή το φάσμα συχνότητας είναι εξαιρετικά μεγάλο 2.5MHz χωρίζεται σε 256 κομμάτια των 10KHz περίπου 340Kbytes δεδομένων αποστέλλονται για επεξεργασία στον χρήστη.

λαμβάνουν τα πακέτα και ο υπολογιστής τους τα επεξεργάζεται και τα αποστέλλει πάλι πίσω στον διακομιστή . Η όλη διαδικασία γίνεται αυτόματα χωρίς να χρειάζεται η συμβολή του

χρήστη και χωρίς να επιβαρύνεται ο προσωπικός υπολογιστής καθώς η όλη επεξεργασία των δεδομένων γίνεται όταν ο υπολογιστής βρίσκεται σε κατάσταση idle. Με αυτόν τον τρόπο αθροίζοντας εκατομμύρια προσκοπικούς υπολογιστές πετυχαίνουν μεγάλη επεξεργαστική ισχύ με πολύ χαμηλό κόστος.[23]

1.3 Computer Cluster

Η επιθυμία για να πάρουμε περισσότερη υπολογιστική ισχύ και για καλύτερη αξιοποίηση της ήδη υπάρχον με χαμηλό κόστος, οδήγησε στην ενορχήστρωση προσωπικών υπολογιστών σε συστοιχίες. Η προσέγγιση του όρου computer clustering συνηθώς (αλλά όχι πάντα) μπορεί να χαρακτηρίσει μια ομάδα υπολογιστών-κόμβων (π.χ προσωπικούς υπολογιστές όπου χρησιμοποιούνται ως Server) μέσω ενός γρήγορου τοπικού δικτύου. Οι δραστηριότητες των υπολογιστών κόμβων ενορχηστρώνονται από το «clustering middleware» ένα λογισμικό όπου βρίσκεται στην «κορυφή» του επιπέδου των κόμβων και διαμορφώνει μια ενιαία εικόνα όλων των κόμβων ως ένα ενιαίο σύστημα. Ένα computer cluster επικεντρώνεται στη διαχείριση όλων των διαθέσιμων υπολογιστών-κόμβων και να δουλεύουν ως ένας server. Είναι διακριτό από άλλες υλοποίησης όπως peer to peer ή grid computing που επίσης χρησιμοποιούν πολλούς υπολογιστές-κόμβους αλλά με πολύ ποιο κατανεμημένη φύση. Οι πρώτες βάσεις για cluster computing ως ένα μέσο για να γίνει παράλληλη εργασία οποιουδήποτε είδους, εφευρέθηκε από τον Gene Amdahl της IBM ο οποίος το 1967 διατύπωσε την θεωρία του με ένα έγγραφο για την παράλληλη επεξεργασία με την ονομασία «Νόμος του Amdahl's.

Εικόνα: 1.4 Beowulf Cluster

Ένα Computer cluster μπορεί να είναι ένα απλό σύστημα το οποίο αποτελείται από δύο μόνο κόμβους οι οποίοι είναι απλοί προσωπικοί υπολογιστές ή μπορεί να είναι μέχρι και ένας πολύ γρήγορος υπερυπολογιστής! Μια βασική προσέγγιση για την οικοδόμηση ενός cluster είναι το Beowulf cluster που μπορεί να κατασκευαστεί με μερικούς προσωπικούς υπολογιστές για να παράγει μια οικονομικά αποδοτική εναλλακτική λύση στα παραδοσιακά High Perfromance Computing (HPC). Ένα από τα πρώτα

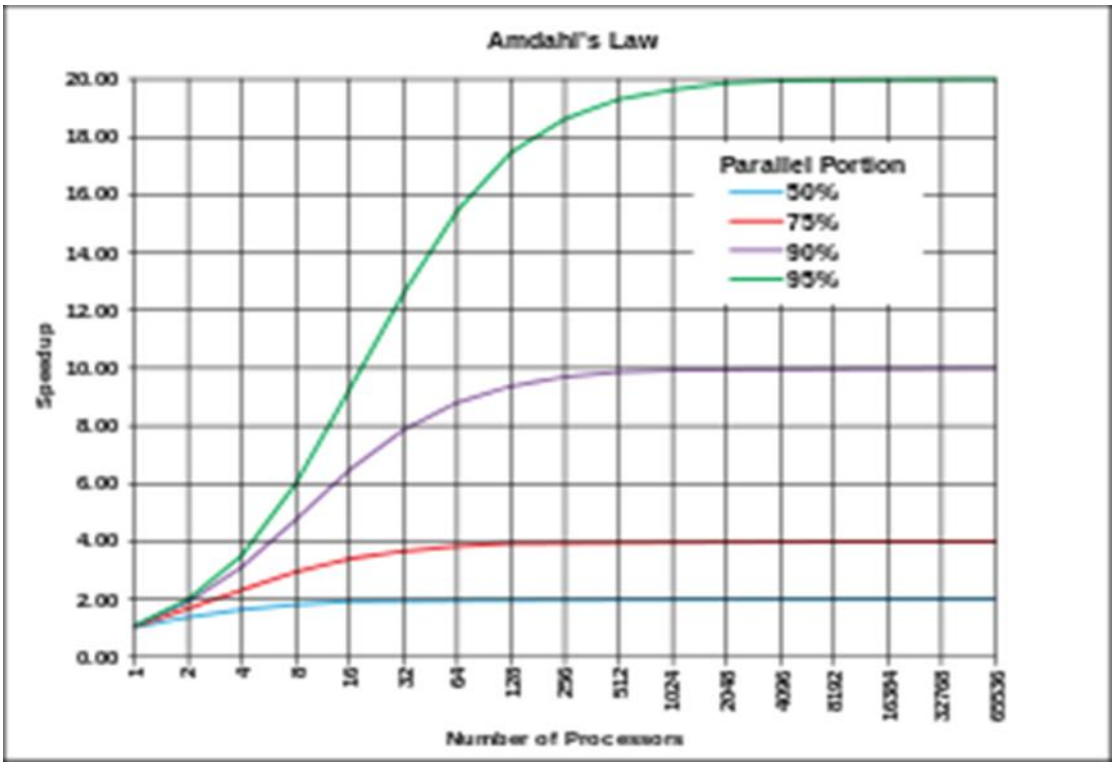
έργα που παρουσίασαν τη βιωσιμότητα της ιδέας ήταν οι 133 κόμβοι Stone Soupercomputer.[24] Οι προγραμματιστές του χρησιμοποίησαν το Linux, την εργαλειοθήκη Parallel Virtual Machine και τη βιβλιοθήκη Interface Message Passing να επιτύχουν υψηλές επιδόσεις σε σχετικά χαμηλό κόστος.[25]

Αν και ένα Cluster μπορεί να αποτελείται από λίγους μόλις προσωπικούς υπολογιστές που συνδέονται με ένα απλό δίκτυο, η αρχιτεκτονική των Cluster μπορεί επίσης να χρησιμοποιηθεί για την επίτευξη πολύ υψηλών επιπέδων απόδοσης. Στην εξαμηνιαία λίστα με τους ταχύτερους υπερυπολογιστές στον κόσμο «TOP500», πολύ συχνά συμπεριλαμβάνονται και πολλά Clusters, για παράδειγμα ο ταχύτερος υπερυπολογιστής στον κόσμο για το 2011 ήταν ο K-computer της Fujitsu όπου βασίζεται στην αρχιτεκτονική Cluster.

1.3.1 Amdahl´s law

Νόμος του Amdahl, επίσης γνωστή ως το επιχείρημα του Amdahl, χρησιμοποιείται για να βρούμε τη μέγιστη αναμενόμενη βελτίωση σε ένα συνολικό σύστημα όταν μόνο μέρος του συστήματος βελτιώνεται. Συχνά χρησιμοποιείται σε parallel computing για να προβλέψει την θεωρητική μέγιστη επιτάχυνση με τη χρήση πολλαπλών επεξεργαστών. Ο νόμος έχει πάρει το όνομά του από τον «αρχιτέκτονα υπολογιστών» Gene Amdahl, και παρουσιάστηκε στο AFIPS Spring στην Διάσκεψη Υπολογιστών το 1967. Η επιτάχυνση ενός προγράμματος με τη χρήση πολλαπλών επεξεργαστών σε παράλληλα συστήματα περιορίζεται από το χρόνο που απαιτείται για τη διαδοχικά κλάσματα-μέρει του προγράμματος. [26]

Για παράδειγμα, εάν ένα πρόγραμμα χρειάζεται 20 ώρες χρησιμοποιώντας ένα μονοπύρηνο επεξεργαστή και ένα συγκεκριμένο τμήμα του προγράμματος που διαρκεί μία ώρα για να εκτελέσει δεν μπορεί να γίνει παραλληλοποίηση , ενώ οι υπόλοιπες 19 ώρες (95%) του χρόνου εκτέλεσης μπορεί να γίνει παραλληλοποίηση , τότε ανεξάρτητα πόσα επεξεργαστές είναι αφιερωμένη σε μια παραλληλισθέντες εκτέλεση του προγράμματος αυτού, ο ελάχιστος χρόνος εκτέλεσης δεν μπορεί να είναι μικρότερος από μία ώρα. Εξ ου και η επιτάχυνση περιορίζεται το πολύ σε 20 ×.



Εικόνα: 1.5 Η επιτάχυνση ενός προγράμματος με τη χρήση πολλαπλών επεξεργαστών. Για παράδειγμα, εάν μπορεί να παραλληλίζεται το 95% του προγράμματος, η θεωρητική μέγιστη επιτάχυνση χρησιμοποιώντας parallel computing θα είναι 20 ×, όπως φαίνεται στο διάγραμμα, δεν έχει σημασία πόσο πολλοί επεξεργαστές χρησιμοποιούνται.

1.3.1.1 Ορισμός

Δίνεται:

- $n \in \mathbb{N}$, ο αριθμός των νημάτων της εκτέλεσης
- $B \in [0,1]$, το κλάσμα του αλγορίθμου που είναι αυστηρά σειριακό
- Ο χρόνος $T(n)$, ένας αλγόριθμος που χρειάζεται για να εκτελεστεί σε n νήμα(τα), η εκτέλεση αντίστοιχη σε :

$$T(n) = T(1)\left(B + \frac{1}{n}(1-B)\right) \quad (1)$$

Ως εκ τούτου, η θεωρητική επιτάχυνση που μπορεί να είχε με την εκτέλεση ενός δεδομένου αλγορίθμου σε ένα σύστημα ικανό να εκτελεί n νήματα της εκτέλεσης είναι:

$$S(n) = \frac{T(1)}{T(n)} = \frac{T(1)}{T(1)\left(B + \frac{1}{n}(1-B)\right)} = \frac{1}{B + \frac{1}{n}(1-B)} \quad (2)$$

Ο Νόμος του Amdahl είναι ένα μοντέλο για τη σχέση μεταξύ της αναμενόμενης επιτάχυνσης της παραλληλισθέντας υλοποιήσεις ενός αλγορίθμου σε σχέση με το σειριακό αλγόριθμο, με βάση την παραδοχή ότι το μέγεθος του προβλήματος παραμένει το ίδιο όταν παραλληλοποιηται. Για παράδειγμα, εάν για ένα δεδομένο μέγεθος του προβλήματος μια παραλληλοποιημένη εφαρμογή ενός αλγορίθμου μπορεί να τρέξει 12% των λειτουργιών του αλγορίθμου αυθαίρετα γρήγορα (ενώ το υπόλοιπο 88% των εργασιών δεν είναι δυνατό να παραλληλιστούν), νόμος του Amdahl δηλώνει ότι η μέγιστη επιτάχυνση του παραλληλισθέντες είναι 1 / (1 - 0,12) = 1.136 φορές τόσο γρήγορα όσο η μη-εφαρμογή παραλληλοποιημένων.

Περισσότερα από τεχνική άποψη, ο νόμος αφορά την εφικτή επιτάχυνση από μια βελτίωση σε έναν υπολογισμό που επηρεάζει ένα ποσοστό P του εν λόγω υπολογισμού, όπου η βελτίωση έχει επιτάχυνση του S. (Για παράδειγμα, αν το 30% του υπολογισμού μπορεί να αποτελέσει αντικείμενο μιας επιτάχυνσης P που θα είναι 0.3. Αν η βελτίωση αυτή καθιστά αυτό το τμήμα δύο φορές πιο γρήγορα, S θα είναι 2.) Ο νόμος του Amdahl αναφέρει ότι η συνολική επιτάχυνση της εφαρμογής θα είναι βελτιωμένη :

$$\frac{1}{(1-P) + \frac{P}{S}} = \frac{1}{(1-0.3) + \frac{0.3}{2}} = 1.1765 \quad (3)$$

Για να δούμε πώς αυτός ο τύπος λειτουργεί, ας υποθέσουμε ότι ο χρόνος εκτέλεσης του παλιού υπολογισμού ήταν 1, για κάποιο μονάδα του χρόνου. Ο χρόνος εκτέλεσης του νέου υπολογισμού θα είναι το χρονικό διάστημα που το «αβελτίωτο» κλάσμα παίρνει (η οποία είναι 1 - P), συν το χρονικό διάστημα για το «βελτιωμένο» κλάσμα παίρνει. Το χρονικό διάστημα για το βελτιωμένο μέρος του υπολογισμού είναι το βελτιωμένης λειτουργίας μήκος χρόνο όπου διαιρείται από την επιτάχυνση, κάνοντας το μήκος του χρόνου του βελτιωμένου τμήματος P / S. Η τελική επιτάχυνση υπολογίζεται διαιρώντας το παλιό χρόνο λειτουργίας από το νέο χρόνο λειτουργίας, το οποίο είναι αυτό που κάνει ο παραπάνω τύπος.[27]

1.3.1.2 Parallel computing

Στην περίπτωση των παράλληλων υπολογισμών, ο νόμος του Amdahl αναφέρει ότι εάν το P είναι το ποσοστό ενός προγράμματος που μπορεί να γίνει παράλληλα (δηλαδή, να επωφεληθούν από τον παραλληλισμό), και (1 - P) είναι το ποσοστό που δεν μπορεί να παραλληλίζεται (παραμένει σειριακό), τότε η μέγιστη επιτάχυνση που μπορεί να επιτευχθεί με τη χρήση επεξεργαστών N είναι :

$$S(N) = \frac{1}{(1-P) + \frac{P}{N}} \quad (4)$$

Στο όριο, καθώς N τείνει στο άπειρο, η μέγιστη επιτάχυνση τείνει να 1 / (1 - P). Στην πράξη, η απόδοση σε αναλογία τιμής πέφτει γρήγορα όσο το N αυξάνεται τη στιγμή που υπάρχει ακόμη και μια μικρή συνιστώσα του (1 - P).

Ως παράδειγμα, εάν το P είναι 90%, τότε (1 - P) είναι 10%, και το πρόβλημα μπορεί να επιταχυνθεί κατά το μέγιστο συντελεστή 10, δεν έχει σημασία πόσο μεγάλη είναι η τιμή του N που χρησιμοποιούμαι. Για το λόγο αυτό, parallel computing είναι χρήσιμο μόνο είτε για μικρό αριθμό επεξεργαστών, ή προβλήματα με πολύ υψηλές τιμές P: Ονομάζοντάς τα παράλληλα προβλήματα. Ένα μεγάλο μέρος της τέχνης του παράλληλου προγραμματισμού αποτελείται από προσπάθειες να μειώσουν το συστατικό (1 - P) με την μικρότερη δυνατή τιμή.

Το P μπορεί να εκτιμηθεί χρησιμοποιώντας την μετρηθείσα επιτάχυνση (SU) σε ένα συγκεκριμένο αριθμό επεξεργαστών (NP) χρησιμοποιώντας:

$$P_{\text{estimated}} = \frac{\frac{1}{SU} - 1}{\frac{1}{NP} - 1} \quad (5)$$

P estimated, με τον τρόπο αυτό μπορεί στη συνέχεια να χρησιμοποιηθεί σε νόμο του Amdahl για να προβλέψουμε επιτάχυνση για ένα διαφορετικό αριθμό επεξεργαστών.

Αναφορές και Βιβλιογραφία

Πρωτεύουσες Αναφορές :

- **[**1] Computer Cluster (n.d.) In Wikipedia. Retrieved 20-08-2013 https://en.wikipedia.org /wiki/Computer_cluster
- [2]History of Supercomputing (n.d) In Wikipedia Retrieved 20-08-2013 https://en.wikipedia.org /wiki/History_of_supercomputing
- [3] Supercomputer (n.d.)In Wikipedia. Retrieved 20-08-2013 https://en.wikipedia.org/wiki /Supercomputer
- [4] Supercomputer (n.d) In Wikipedia Retrieved 21-08-2013 https://en.wikipedia.org/wiki /Supercomputer
- [5]"Sublist Generator". top500.org. Retrieved 21-08-2013. http://www.top500.org/blog/lists /2012/11/press-release/
- [6] TOP500 Annual Report 1994. Retrieved 21-08-2013 http://www.intel.com/content/dam /doc/report/history-1994-annual-report.pdf
- [7] Numerical Wind Tunnel(n.d) In Wikipedia Retrieved 22-08-2013 https://en.wikipedia.org /wiki/Numerical_Wind_Tunnel_(Japan)
- [8]Tiahne-2(n.d) In Wikipedia. Retrieved 17-06-2014 .https://en.wikipedia.org/wiki/Tianhe-2
- [9]: Tianhe-2(n.d)In Wikipedia. Retrieved 2014-07-10. https://en.wikipedia.org/wiki/Tianhe-2
- [10]Tiahne-2 Michael Kan, IDG News Service (2012-10-31). "China is building a 100-petaflop supercomputer". infoworld.com. Retrieved 2012-10-31
- [11] Tianhe-2(n.d) In Wikipedia.Retrieved 24-06-2014. https://en.wikipedia.org/wiki/Tianhe-2
- [12] Tianhe-2(n.d) In Wikipedia.Retrieved 23 -08- 2014. https://en.wikipedia.org/wiki/Tianhe-2
- [13] Supercomputer (n.d) In Wikipedia Retrieved 24-08-2014. https://en.wikipedia.org/wiki /Supercomputer
- [14]Supercomputer (n.d)In Wikipedia Retrieved 17-06-2014. https://en.wikipedia.org/ wiki /Supercomputer

  [15] Supercomputer (n.d) In Wikipedia Retrieved 19-06-2014 https://en.wikipedia.org/wiki/ Supercomputer,https://upload.wikimedia.org/wikiversity/en/0/00/Cloud_Computing wiki. 20150310.pdf
- [16]Supercomputer (n.d)In Wikipedia . Retrieved 31-07-2014 . https://en.wikipedia.org/wiki/Supercomputer

- [17] Grid Computing (n.d) In Wikipedia. Retrieved 18-09-2013. https://en Wikipedia.org/wiki /Grid_computing
- [18] Supercomputer (n.d) In Wikipedia. Retrieved 18-09-2013. https://en.wikipedia.org/wiki /Supercomputer
- [19] Grid Computing (n.d) In Wikipedia. Retrieved 20-08-2014. https://en Wikipedia.org/wiki /Grid_computing
- [20] BOINC, Choosing BOINC project, Retrieved 22-08-2014. https://boinc.berkeley.edu /projects.php
- [21] BOINC, Other sources of BOINC client software, Retrieved 22-08-2014 https://boinc . berkeley.edu/trac/wiki/DownloadOther
- [22] Seti@home (n.d) In Wikipedia. Retrieved 23-08-2014, https://en.wikipedia.org/wiki/ SETI@home
- [23] seti.berkeley.edu Retrieved 20-06-2014
- [24]Computer Cluster Network (n.d) In Wikipedia. Retrieved 20-06-2014 https://en.wikipedia .org/wiki/Computer_cluster
- [25] Computer Cluster (16-08-2001) Retrieved 18-10-2013. https://en.wikipedia.org/wiki /Computer_cluster
- [26]Amdahl´s Law (n.d) In Wikipedia. Retrieved 18-10-2013. https://en.wikipedia.org/wiki /Amdahl%27s_law
- [27] Amdahl´s Law (n.d) In Wikipedia Retrieved 24-10-2013. https://en.wikipedia.org/wiki /Amdahl%27s_law.

Δευτερεύουσες Αναφορές :

- Computer cluster, Readings in computer architecture by Mark Donald Hill, Norman Paul Jouppi, Gurindar Sohi 1999 ISBN 978-1-55860-539-8 page 41-48, https://en.wikipedia.org /wiki/Computer_cluster
- History of Supercomputing, Milestones in computer science and information technology by Edwin D. Reilly 2003 ISBN 1-57356-521-0 page 65.https://en.wikipedia.org /wiki/History_of_supercomputing
- Supercomputer, Hoffman, Allan R.; et al. (1990). Supercomputers: directions in technology and applications. National Academies. pp. 35–47. ISBN 0-309-04088-4. https://en.wikipedia.org /wiki/Supercomputer
- Supercomputer, Hill, Mark Donald; Jouppi, Norman Paul; Sohi, Gurindar (1999). Readings in computer architecture. pp. 40–49. ISBN 1-55860-539-8.https://en.wikipedia.org/wiki /Supercomputer.

- Numerical Wind Tunnel N. Hirose and M. Fukuda (1997). "Numerical Wind Tunnel (NWT) and CFD Research at National Aerospace Laboratory". Proceedings of HPC-Asia '97. IEEE Computer Society. https://en.wikipedia.org /wiki/Numerical_Wind_Tunnel_(Japan)
- Tiahne-2,"June2013".TOP500. https://en.wikipedia.org/wiki/Tianhe-2
- Tiahne-2"TheTop 500 List: june 2013. https://en.wikipedia.org/wiki/Tianhe-2
- Tiahne-2"China's Tianhe-2 Remains The World's Fastest Supercomputer". Forbes. https://en.wikipedia.org/wiki/Tianhe-2
- Tiahne-2 "China's Tianhe-2 Remains The World's Fastest Supercomputer". Forbes. 23 August 2014. https://en.wikipedia.org/wiki/Tianhe-2
- Supercomputer,"Internet PrimeNet Server Distributed Computing Technology for the Great Internet Mersenne Prime Search". GIMPS. Retrieved 6 June 2011. https://en.wikipedia.org/wiki/Supercomputer
- Supercomputer,"Folding@home: OS Statistics". Stanford University https://en.wikipedia.org/wiki/Supercomputer
- Supercomputer, Cloud computing"BOINCstats: BOINC Combined". BOINC. Retrieved 28 May 2011Note this link will give current statistics, not those on the date last accessed https://en.wikipedia.org/wiki/Supercomputer
- Supecomputer, "SETI@Home Credit overview". BOINC https://en.wikipedia.org/wiki /Supercomputer
- Grid Computing, What is grid computing? - Gridcafe. E-sciencecity.org https://en Wikipedia.org/wiki/Grid_computing
- Grid Computing,"A Gentle Introduction to Grid Computing and Technologies" .. Buyya, Rajkumar; Kris Bubendorfer (2009). Market Oriented Grid and Utility Computing. Wiley. ISBN 978-0-470-28768-2 https://en.wikipedia.org/wiki/Supercomputer
- Grid computing, Computational problems - Gridcafe. E-sciencecity.org https://en Wikipedia.org/wiki/Grid_computing
- Seti@home, Porting and optimizing SETI@home https://en.wikipedia.org/wiki
- /SETI@home
- Computer cluster,Network-Based Information Systems: First International Conference, NBIS 2007 ISBN 3-540-74572-6 page 375 https://en.wikipedia.org/wiki/Computer_cluster
- Computer cluster,William W. Hargrove, Forrest M. Hoffman and Thomas Sterling (August 16, 2001). "The Do-It-Yourself Supercomputer". Scientific American 265 (2). pp. 72–79. https://en.wikipedia.org/wiki/Computer_cluster
- Amdahl`s law ,Rodgers, David P. (June 1985). "Improvements in multiprocessor system design". ACM SIGARCH Computer Architecture News archive (New York, NY, USA: ACM) 13 https://en.wikipedia.org/wiki /Amdahl%27s_law
- Amdahl`s law, Amdahl, Gene M. (1967). "Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities". AFIPS Conference Proceedings. https://en.wikipedia.org/wiki/Amdahl%27s_law.

ΚΕΦΑΛΑΙΟ 2

Computer Cluster

2.1 Η Πρώτη εμφάνιση Computer Cluster

Σύμφωνα με των Greg Pfister που έχει γράψει σχετικά για τα Clusters (¨In Search of Clusters¨), τα clusters δεν εφευρέθηκαν από κάποιον συγκεκριμένο προμηθευτή αλλά από τους ίδιους τους πελάτες και τις ανάγκες τους, που δεν μπορούσαν να χωρέσουν όλες τις εργασίες τους σε έναν υπολογιστή ή χρειάζονταν δυνατότητα back-up με μικρό κόστος.[1] Οι πρώτες όμως βάσεις για για cluster computing ως ένα μέσο για να γίνει παράλληλη εργασία οποιουδήποτε είδους, αναμφισβήτητα εφευρέθηκε από τον Gene Amdahl της IBM ο οποίος το 1967 με ένα έγγραφο για την παράλληλη επεξεργασία με την ονομασία «Νόμος του Amdahl's. Η ιστορία των πρώτον Clusters είναι συνδεδεμένη και με την ανάπτυξη των πρώτον δικτύων, λειτούργησε ως ένα επιπλέων κίνητρο στο να συνδέσουν του υπολογιστικούς πόρους μεταξύ τους, δημιουργώντας έτσι τα πρώτα Clusters. Το πρώτο εμπορικό προϊόν cluster ήταν το Arcnet, που αναπτύχθηκε από την Datapoint το 1977.

Αλλά το πρώτο εμπορικά επιτυχημένο Cluster ήταν της Digital Equipment Corporation που κυκλοφόρησε το προϊόν τους VAXcluster το 1984 με το λειτουργικό σύστημα VAX / VMS.

Τα προϊόντα ARCNET και VAXcluster υποστηρίζαν όχι μόνο parallel computing, αλλά και κοινά συστήματα αρχείων και κοινή χρήση των περιφερειακών συσκευών. Η ιδέα ήταν να προσφέρει τα πλεονεκτήματα της παράλληλης επεξεργασίας και τη διατήρηση της αξιοπιστίας των δεδομένων.

Το ίδιο χρονικό διάστημα άρχισαν να αξιοποιούν την παραπάνω τεχνολογία και υπερυπολογιστές, χαρακτηριστικό παράδειγμα ο Cray-1 το 1976 ο οποίος εισήγαγε εσωτερική παράλληλη επεξεργασία δεδομένων.[2] Στα χρόνια που ακολούθησαν οι υπερυπολογιστές έκαναν εκτενέστερη χρήση του clustering πλέον και με εξωτερικές δικτυώσεις, έχοντας φτάσει σήμερα σε ταχύτατους υπερυπολογιστές όπου κάνουν χρήστη του Clustering, όπως είναι για παράδειγμα ο K-Computer της Fujitsu.

Εικόνα: 2.1 K-Computer Fujitsu χαρακτηριστικό παράδειγμα χρήσης Clustering σε υπερυπολγιστές.

2.2 Είδη συστοιχιών Computer Cluster

Computer Clusters μπορούν να ρυθμιστούν για διαφορετικούς σκοπούς, από γενικής χρήσεως επιχειρηματικές ανάγκες όπως υποστήριξη υπηρεσιών διαδικτύου (web-services), μέχρι υποστήριξη για επιστημονικούς υπολογισμούς. Κάθε συστοιχία έχει τα δικά της χαρακτηριστικά, πλεονεκτήματα και μειονεκτήματα. Τα είδη αυτά είναι :

- Συστοιχίες Υψηλής Διαθεσιμότητας (High Availability Clusters)
- Συστοιχίες Εξισορρόπησης Φορτίου (Load Balancing Clusters)
- Συστοιχίες Υψηλής Απόδοσης (High Performance Clusters)

2.2.1 Συστοιχίες Υψηλής Διαθεσιμότητας (High Availability Clusters)

Clusters Υψηλής Διαθεσιμότητας (επίσης γνωστά ως Clusters HA ή Clusters ανακατεύθυνσης) Οι συστοιχίες αυτές έχουν σχεδιαστεί για να προσφέρουν συνεχή πρόσβαση σε εφαρμογές παροχής υπηρεσιών. Διατηρούν επιπλέον κόμβους που μπορούν να χρησιμοποιηθούν σαν εφεδρικά συστήματα στην περίπτωση αστοχίας των κύριων κόμβων. Ο ελάχιστος αριθμός κόμβων σε μία τέτοια συστοιχία είναι δύο (ένας κύριος και ένας εφεδρικός), παρόλο που η συντριπτική πλειοψηφία χρησιμοποιεί περισσότερους κόμβους. Για Λειτουργικό Σύστημα συνηθίζουν να χρησιμοποιούν ελεύθερες διανομές Linux με ανάλογες «σουίτες» για clustering. Τα HA clusters χρησιμοποιούνται συχνά για κρίσιμες βάσεις δεδομένων, την κοινή

χρήση αρχείων στο δίκτυο, επιχειρηματικές εφαρμογές, και την εξυπηρέτηση πελατών, όπως το ηλεκτρονικό εμπόριο-ιστοσελίδες.

Τα HA clusters χρησιμοποιούν συνήθως ένα παλμό-σήμα με την χρήση ιδιωτικού δικτύου για την παρακολούθηση της υγείας και την κατάσταση κάθε κόμβου του Cluster. Σε μια σοβαρή κατάσταση όλο το λογισμικό του cluster πρέπει να είναι σε θέση να χειριστεί ένα split-brain, το οποίο συμβαίνει όταν το σύνολο των συνδέσεων του δικτύου πέσει ταυτόχρονα, αλλά οι κόμβοι του Cluster βρίσκονται σε λιτουργεία και εκτελούν κάποια εργασία. Αν συμβεί αυτό, κάθε κόμβος του Cluster μπορεί εσφαλμένα να αποφασίσει ότι άλλος κόμβος έχει «πέσει» και προσπαθεί να ξεκινήσει τις υπηρεσίες που άλλοι κόμβοι ακόμα επεξεργάζονται. Έχοντας έτσι διπλές παρουσίες των υπηρεσιών μπορεί να προκαλέσει καταστροφή των δεδομένων σχετικά με το κοινόχρηστο αποθηκευτικό χώρο.

2.2.1.1 Απαιτήσεις σχεδιασμού εφαρμογών

Δεν μπορεί κάθε εφαρμογή να τρέξει σε ένα υψηλής διαθεσιμότητας περιβάλλον Cluster, καθώς και οι αναγκαίες αποφάσεις για το σχεδιασμό πρέπει να γίνονται νωρίς στη φάση του σχεδιασμού του λογισμικού. Για να εκτελεστεί μια εφαρμογή σε ένα υψηλής διαθεσιμότητας περιβάλλον Cluster, η αίτηση πρέπει να πληροί τουλάχιστον τις ακόλουθες τεχνικές απαιτήσεις, οι τελευταίες εκ των οποίων δύο είναι ζωτικής σημασίας για την αξιόπιστη λειτουργία του σε ένα Cluster και είναι το πιο δύσκολο να ικανοποιηθούν πλήρως:

- Πρέπει να υπάρχει ένας σχετικά εύκολος τρόπος για να ξεκινήσει, να σταματήσει, δυναμικό σταμάτημα, και να ελέγχετε η κατάσταση της εφαρμογής. Σε πρακτικούς όρους, αυτό σημαίνει ότι η αίτηση πρέπει να έχει ένα περιβάλλον γραμμής εντολών ή χρήση κάποιων scripts για τον έλεγχο της εφαρμογής, συμπεριλαμβανομένης της υποστήριξης για πολλαπλές εμφανίσεις της εφαρμογής.

- Η εφαρμογή πρέπει να είναι σε θέση να χρησιμοποιεί και να μοιράζεται μέσα αποθήκευσης (NAS/SAN).

- Το πιο σημαντικό είναι ότι η εφαρμογή θα πρέπει να αποθηκεύει την κάθε κατάσταση σε ένα μη πτητικό μέσο αποθήκευσης όσον αυτό είναι δυνατόν. Εξίσου σημαντική είναι η δυνατότητα

να επανεκκίνηση σε ένα άλλο κόμβο στην τελευταία κατάσταση πριν από τη βλάβη χρησιμοποιώντας την αποθηκευμένη κατάσταση από τον κοινόχρηστο αποθηκευτικό χώρο.

- Η εφαρμογή δεν πρέπει να καταστρέφει τα δεδομένα εάν για κάποιο λόγο «κολλήσει» , ή γίνει επανεκκίνηση από την υποθηκευμένη κατάσταση.

2.2.1.2 Διαμόρφωση κόμβων

Η πιο κοινή διαμόρφωση για ένα HA είναι ένα Cluster δύο κόμβων, δεδομένου ότι αυτό είναι το ελάχιστο που απαιτείται για την παροχή πλεονασματικού κόμβου-(server), αλλά πολλά Cluster αποτελούνται από πολλούς περισσότερους πλεονασματικούς κόμβους , μερικές φορές ακόμα και από δεκάδες κόμβους. Τέτοιες διατάξεις μπορούν μερικές φορές να ταξινομηθούν σε ένα από τα παρακάτω μοντέλα:

- Ενεργός/Παθητικός (Active/Passive)- Παρέχει για κάθε κόμβο του συστήματος ακόμα έναν έξτρα, παθητικό, ο οποίος ενεργοποιείτε και συνδέεται απευθείας μόνο όταν ο συναφών κύριος κόμβος του αποτύχει. Η διαμόρφωση αυτή απαιτεί συνήθως πολύ επιπλέον υλικό.[3]

- N +1 - Παρέχει έναν επιπλέον κόμβο που φέρει απευθείας σύνδεση για να αναλάβει το ρόλο του κόμβου που έχει αποτύχει. Στην περίπτωση ετερογενών διαμορφώσεων του λογισμικού σε κάθε κύριο κόμβο, ο κόμβος πρέπει επιπλέον να είναι σε θέση να αναλάβει καθολικά οποιοδήποτε από τους ρόλους των πρωτογενών κόμβων για τον οποίο είναι υπεύθυνος. Αυτό αναφέρεται συνήθως σε Clusters τα οποία έχουν πολλαπλές υπηρεσίες που εκτελούνται ταυτόχρονα. Στη περίπτωση μεμονομένης παροχής υπηρεσιών, το μοντέλο γίνεται παρόμοιο με το μοντέλο Ενεργός / Παθητικός (Active/Passive) κόμβος.

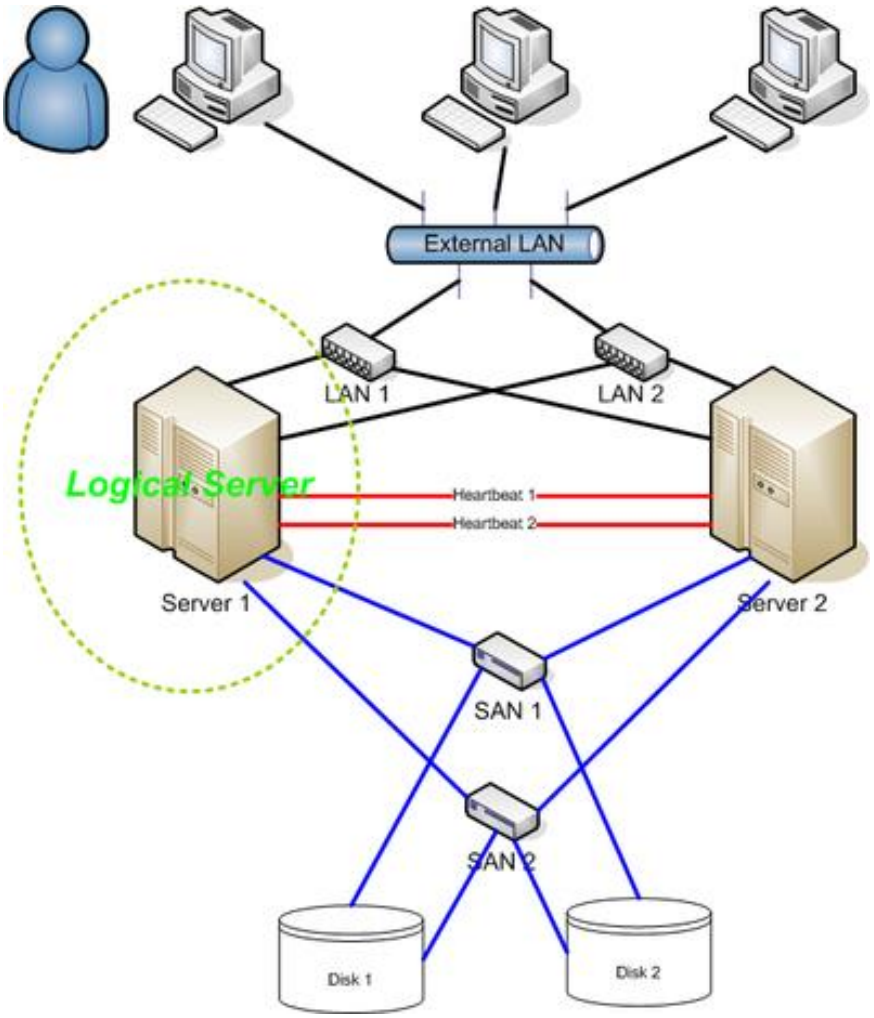
- N + M – Στις περιπτώσεις όπου ένα απλό Cluster διαχειρίζεται πολλές υπηρεσίες, έχοντας μόνο ένα αφιερωμένο κόμβο ανακατεύθυνσης δεν μπορεί να προσφέρει επαρκές πλεόνασμα. Σε τέτοιες περιπτώσεις συμπεριλαμβάνονται, περισσότεροι από έναν (M) standby servers (σε αναμονή) και είναι διαθέσιμοι. Ο αριθμός των standby servers (σε αναμονή) είναι μια ισορροπία μεταξύ του κόστους και των απαιτήσεων αξιοπιστίας.

- N-to-1 - Επιτρέπει ο κόμβος αναμονής να γίνει ενεργός προσωρινά, έως ότου ο αρχικός κόμβος να μπορεί να αποκατασταθεί ή να επανέλθει σε κανονική λειτουργεία. Σε ποιο σημείο των υπηρεσιών ή των περιπτώσεων πρέπει να επανέλθει στο σημείο αυτό και να συνεχίσει τις εργασίες του, ούτως ώστε να αποκατασταθεί η υψηλή διαθεσιμότητα .

- N-to-N - Ο συνδυασμός των ενεργών / ενεργών (active/active) και N + M clusters. Τα N-to-N clusters αναδιανέμουν τις υπηρεσίες, κάνουν υποδείξεις ή τις συνδέσεις στην θέση του κόμβου που έχει πρόβλημα μεταξύ των υπόλοιπων ενεργών κόμβων, εξαλείφοντας έτσι (όπως και με active / active) την ανάγκη για μια «κατάσταση αναμονής» (standby mode) κόμβου, αλλά εισάγοντας την ανάγκη για επιπλέον χωρητικότητα σε όλους τους ενεργούς κόμβους.

Ο όρος Logical host ή Cluster logical host χρησιμοποιείται για να περιγράψει τη διεύθυνση του δικτύου που χρησιμοποιείται για την πρόσβαση στις υπηρεσίες που παρέχονται από το Cluster. Η ταυτότητα του Logical Host δεν είναι συνδεδεμένη με έναν κόμβο του Cluster. Είναι στην πραγματικότητα μία διεύθυνση δικτύου/ hostname που συνδέεται με την υπηρεσία(ες) που παρέχεται από το Cluster.

Εάν ένας κόμβος του Cluster με μια βάση δεδομένων που τρέχει, ξαφνικά παρουσιάσει πρόβλημα και «πέσει», η βάση δεδομένων θα πρέπει να είναι σε θέση να ξαναρχίσει σε άλλον κόμβο του Cluster, και η διεύθυνση του δικτύου που οι χρήστες χρησιμοποιούν για να έχουν πρόσβαση στη βάση δεδομένων, θα μεταφερθεί για το νέο κόμβο, έτσι ώστε οι χρήστες να μπορούν έχουν πρόσβαση στη βάση δεδομένων πάλι.



Σχήμα: 2.1 Διάγραμμα δικτύου κόμβων High Availability Cluster

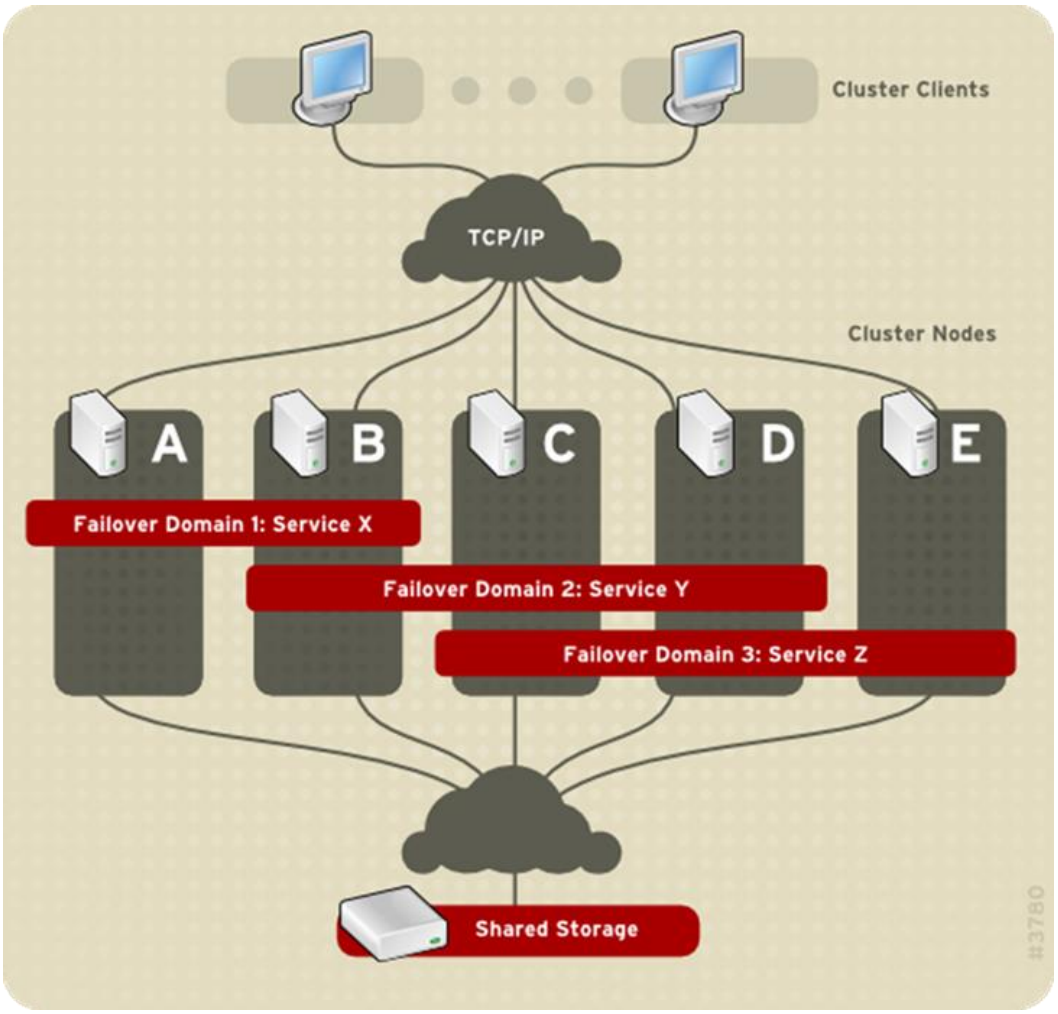
2.2.1.3 Αξιοπιστία κόμβων

Τα HA clusters συνήθως χρησιμοποιούν όλες τις διαθέσιμες τεχνικές για να κάνουν τα επιμέρους συστήματα και την κοινή υποδομή όσο το δυνατόν πιο αξιόπιστο. Αυτά περιλαμβάνουν:

- Disk mirroring στην περίπτωση αστοχίας των εσωτερικών δίσκων δεν οδηγεί σε κατάρρευση του συστήματος. Η Distributed Device Replicated Block είναι ένα παράδειγμα.

- Επιπλέον Συνδέσεις δικτύου (Redundant network connections) έτσι ώστε αστοχίες μονού καλωδίου, του switch, ή αποτυχίες διασύνδεσης δικτύου δεν οδηγούν σε διακοπές λειτουργίας του δικτύου.

- Εφεδρικά συστήματα για εισροή ηλεκτρικής ενέργειας με διαφορετικά κυκλώματα, συνήθως και οι δύο κεντρικοί κόμβοι ή όλοι προστατεύονται από αδιάλειπτης παροχής ηλεκτρικής ενέργειας και οι μονάδες εφεδρική τροφοδοσία, έτσι ώστε σε περίπτωση αστοχίας μια γεννήτριας τροφοδοσίας , καλωδίου ή UPS δεν οδηγούν σε απώλεια ισχύος όλο το σύστημα.



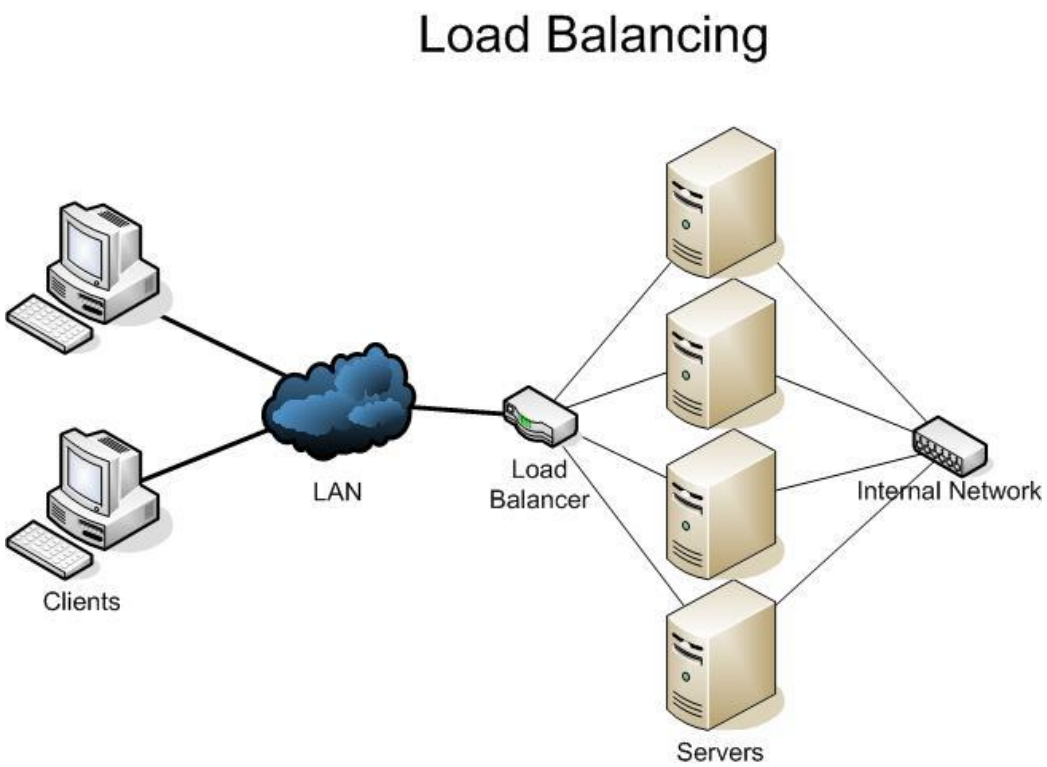
Σχήμα: 2.2 Χαρακτηριστικό High Aveilability Cluster με χρήση υπηρεσίας ανακατεύθυνσεις (failover)

Αυτά τα χαρακτηριστικά βοηθούν να ελαχιστοποιηθούν οι πιθανότητες ότι θα απαιτηθεί η ανακατεύθυνση του Cluster μεταξύ των συστημάτων. Σε μια τέτοια περίπτωση, η

παρεχόμενη υπηρεσία ανακατεύθυνσης δεν είναι διαθέσιμη για λίγο, έτσι ώστε να προτιμηθούν άλλα μέτρα για την αποφυγή ανακατεύθυνσης (failover).

2.2.2 Συστοιχίες Κατανομής Φορτίου (Load Balancing Clusters)

Οι συστοιχίες αυτές ονομάζονται Κατανομής Φορτίου ή Εξισορρόπησης Φορτίου ή Load Balancin Clusters. Οι συστοιχίες αυτού του τύπου, είναι μια μέθοδος δικτύωσης υπολογιστών για τη διανομή του φόρτου εργασίας σε πολλούς υπολογιστικούς πόρους όπως υπολογιστές, υπολογιστές Cluster , συνδέσεις δικτύου, κεντρικές μονάδες επεξεργασίας ή δίσκους. Οι Συστοιχίες Κατανομής φορτίου στοχεύουν στη βελτιστοποίηση της χρήσης των πόρων, τη μεγιστοποίηση της απόδοσης, την ελαχιστοποίηση του χρόνου απόκρισης, και στην αποφυγή της υπερφόρτωσης του κάθε ενός από τους πόρους.



Σχήμα: 2.3 Χαρακτηριστική συστοιχία Cluster Κατανομής Φορτίου (Load Balancing). Διακρίνετε οι πελάτες (clients)να συνδέονται μέσω ενός εξωτερικού δικτύου (μπορεί να είναι ένα τοπικό δίκτυο LAN ή ακομα και σύνδεση internet) σε μία συσκευή Κατανομής Φορτίου (Load Balancer) και αυτή να συνδέεται σε Servers όπου με τη σειρά τους συνδέονται σε ένα εσωτερικό δίκτυο (internal network), όπου μπορούν να υπάρχουν nodes για την υποστήριξη των servers , βάσεις δεδομένων κλπ., ανάλογα με τη χρήση του cluster.

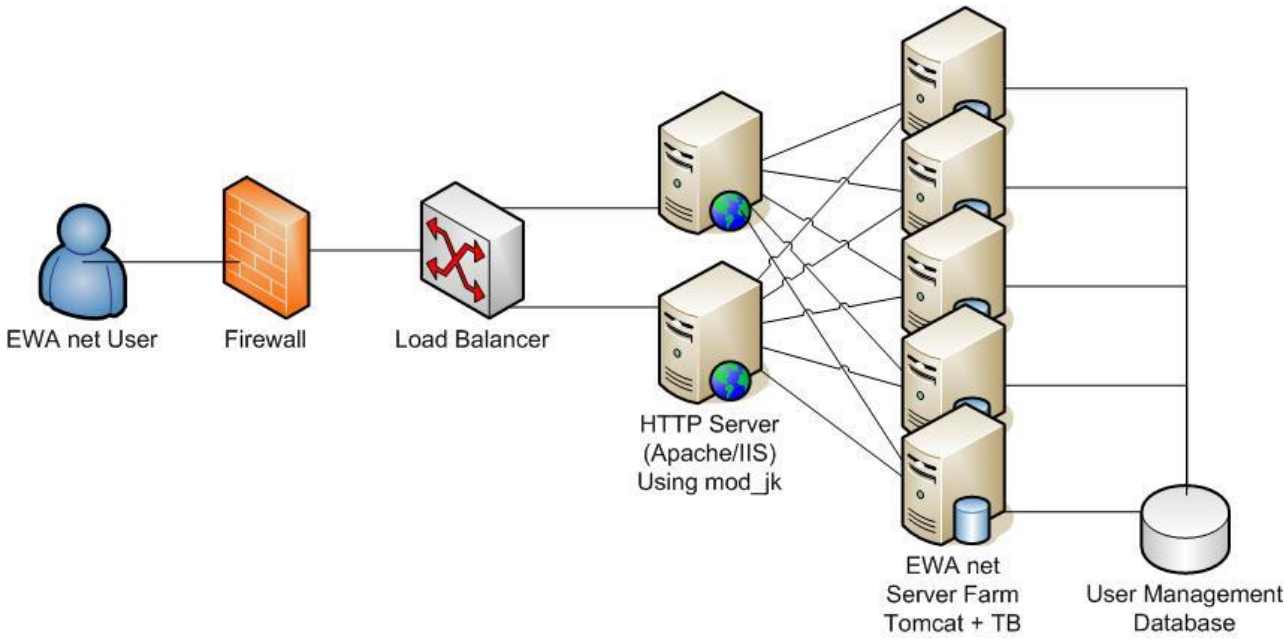
Για το σκοπό αυτό μεταφέρουν διεργασίες από τον ένα κόμβο στον άλλο, ανάλογα με τον φόρτο που έχει το κάθε σύστημα. Χρησιμοποιώντας πολλούς πόρους με εξισορρόπηση

φορτίου, αντί ενός ενιαίου πόρου μπορεί να αυξηθεί η αξιοπιστία μέσω του πλεονασμού των πόρων. Η Συστοιχία Κατανομής φορτίου, συνήθως, παρέχεται από ειδικό λογισμικό ή υλικό, όπως ένα multilayer switch ή μια διεργασία DNS Server. Οι συστοιχίες εξισορρόπησης φορτίου είναι ιδιαίτερα χρήσιμες σε αυτούς που δουλεύουν με περιορισμένο προϋπολογισμό, γιατί φροντίζουν για την όσο το δυνατό αποδοτικότερη εκμετάλλευση του υπάρχοντος εξοπλισμού.

2.2.2.1 Χρήσεις Συστοιχίας Κατανομής Φορτίου

Μία από τις πιο συχνά χρησιμοποιούμενες εφαρμογές της κατανομής φορτίου είναι να παρέχει μια ενιαία υπηρεσία Internet από πολλούς διακομιστές (servers), μερικές φορές γνωστοί ως server farm. Συνήθως τα συστήματα εξισορρόπηση φορτίου περιλαμβάνουν δημοφιλείς ιστοσελίδες, μεγάλα δίκτυα Internet Relay Chat, υψηλού εύρους ζώνης ιστοσελίδες File Transfer Protocol , Network News Transfer Protocol (NNTP) servers και servers Domain Name System (DNS). Τον τελευταίο καιρό, κάποιες συστοιχίες κατανομής φορτίου έχουν εξελιχθεί για να υποστηρίζουν βάσεις δεδομένων. Αυτές οι συστοιχίες ονομάζονται συστοιχίες κατανομής φορτίου βάσεων δεδομένων (Database Load Balancers).

Για τις υπηρεσίες Internet, ο Load Balancer είναι συνήθως ένα πρόγραμμα λογισμικού που εκτελεί ακρόαση στην είσοδο όπου οι εξωτερικοί πελάτες (clients) συνδέονται στις υπηρεσίες πρόσβασης. Το σύστημα εξισορρόπησης φορτίου στέλνει forward request σε ένα από τα "backend" servers, τα οποία συνήθως απαντούν στο σύστημα εξισορρόπησης φορτίου. Αυτό επιτρέπει στο σύστημα εξισορρόπησης φορτίου να απαντήσει στον πελάτη, χωρίς ο πελάτης(client) ποτέ να μάθει για την δομή και για τον εσωτερικό διαχωρισμό των λειτουργιών. Επίσης, αποτρέπει τους πελάτες από την απευθείας επικοινωνία με back-end servers, το οποίο μπορεί να έχει οφέλη για την ασφάλεια, αποκρύπτοντας τη δομή του εσωτερικού δικτύου και αποτρέπει τυχών επιθέσεις στον πυρήνα της στοιβάδας του δικτύου ή άσχετες υπηρεσίες που εκτελούνται σε άλλες εισόδους.



Σχήμα: 2.4 Χρήση συστοιχίας κατανομής φορτίου με βάση δεδομένων. Ο χρήστης συνδέεται στο Firewall όπου είναι ένα σύνολο κανόνων που αποφασίζουν εάν η κίνηση μπορεί να περάσει μέσω μιας διεπαφής ή όχι και προστατεύει από κακόβουλους χρήστες, στην συνέχεια ο Load Balancer παρέχει μια ενιαία υπηρεσία internet στους servers HTTP servers, οι οποίοι με τη σειρά τους συνδέονται σε πολλούς servers γνωστοί και ως server farm και όλοι μαζί μοιράζονται μια κοινή βάση δεδομένων.

Μερικοί Load Balancers παρέχουν ένα μηχανισμό για να κάνει κάτι ιδιαίτερο στην περίπτωση που όλοι οι servers backend δεν είναι διαθέσιμοι. Αυτό μπορεί να περιλαμβάνει τη διαβίβαση σε ένα buckup Load Balancer ή να εμφανίζει ένα μήνυμα σχετικά με τη διακοπή. Το σύστημα Εξισορρόπησης φορτίου δίνει στην ομάδα IT την ευκαιρία να επιτύχει μια αισθητά μεγαλύτερη ανοχή σφαλμάτων. Μπορεί να παρέχει αυτόματα το ποσό της χωρητικότητας που απαιτείται για να ανταποκριθεί σε οποιαδήποτε αύξηση ή μείωση αίτησης κυκλοφορίας.

Είναι επίσης σημαντικό το γεγονός ότι ο ίδιος ο Load Balancer δεν θα πρέπει να γίνει ποτέ ένα σημείο αποτυχίας (point of failure). Συνήθως οι Load Balancers εφαρμόζονται σε υψηλής διαθεσιμότητας ζεύγη που μπορούν επίσης να αναπαράγουν τα στοιχεία συνόδου, εφόσον απαιτείται από τη συγκεκριμένη εφαρμογή.[4]

2.2.2.2 Round-robin DNS

Μία εναλλακτική μέθοδος κατανομής φορτίου, η οποία δεν απαιτεί κατ' ανάγκη ένα ειδικό λογισμικό ή hardware κόμβο (node), καλείται round robin DNS. Σε αυτήν την τεχνική, πολλαπλές διευθύνσεις IP συνδέονται με ένα μόνο domain name. Οι πελάτες περιμένουν να

επιλέξουν έναν server για να συνδεθούν. Σε αντίθεση με την χρήση ενός ξεχωριστού κατανεμητή φορτίου, η τεχνική αυτή εκθέτει στους πελάτες την ύπαρξη πολλαπλών backend servers. Η τεχνική έχει και άλλα πλεονεκτήματα και μειονεκτήματα, ανάλογα με το βαθμό ελέγχου επί του DNS server και την επιθυμία για διασπορά της κατανομής φορτίου.[5]

Μια άλλη πιο αποτελεσματική τεχνική για την Κατανομή Φορτίου χρησιμοποιώντας DNS είναι να αναθέσει την www.example.org ως sub-domain του οποίου η ζώνη εξυπηρετείται από τον καθένα από τους ίδιους servers που εξυπηρετούν την ιστοσελίδα. Αυτή η τεχνική λειτουργεί ιδιαίτερα καλά όταν οι επιμέρους servers είναι διασκορπισμένοι γεωγραφικά στο Διαδίκτυο. Για παράδειγμα,

```
one.example.org A 192.0.2.1
two.example.org A 203.0.113.2
www.example.org NS one.example.org
www.example.org NS two.example.org
```

Ωστόσο, το αρχείο ζώνης για www.example.org σε κάθε server είναι διαφορετικό, έτσι ώστε κάθε server επιλύει δική του διεύθυνση IP όπως ο A-record. Στον Server « Ένα» το αρχείο ζώνης για www.example.org αναφέρεται:

```
Στην 192.0.2.1
```

Στον Server «Δύο» το αρχείο ζώνης περιλαμβάνει:

```
Την 203.0.113.2
```

Με αυτό τον τρόπο, όταν ένας server είναι «πέσει», το DNS του δεν θα ανταποκρίνεται και η διαδικτυακή υπηρεσία δεν θα λάβει καμία κίνηση. Εάν η γραμμή σε ένα server είναι κορεσμένη, η αναξιοπιστία του DNS εξασφαλίζει να φτάνει λιγότερη κίνηση HTTP σε αυτόν το server. Επιπλέον, ο πιο γρήγορος DNS απαντάει στο αναλυτή που είναι σχεδόν πάντα ένας από το πιο κοντινούς εξυπηρετητές (servers) του δικτύου, διασφαλίζοντας γεω-ευαίσθητη

Κατανομή Φορτίου. Μια σύντομη TTL για το A-record βοηθά στο να εξασφαλιστεί ότι η κυκλοφορία εκτρέπεται γρήγορα όταν ένας κεντρικός υπολογιστής πηγαίνει να «πέσει». Πρέπει να εξετασθεί το ενδεχόμενο ότι αυτή η τεχνική μπορεί να προκαλέσει μεμονωμένους πελάτες (clients) όπου εναλλάσσονται μεταξύ των ξεχωριστών servers στα μέσα μίας συνεδρίας.

2.2.2.3 Αλγόριθμοι Προγραμματισμού

Πολυάριθμοι αλγόριθμοι προγραμματισμού χρησιμοποιούνται από Κατανεμητές Φορτίου για να καθορίσουν σε ποιον back-end server να στείλουν αίτημα. Απλοί αλγόριθμοι περιλαμβάνουν τυχαία επιλογή ή round robin. Πιο εξελιγμένοι Κατανεμητές φορτίου μπορεί να λάβουν επιπλέον παράγοντες υπόψη, όπως την αναφορά φορτίου των server, επιπλέων χρόνους απόκρισης, την κατάσταση λειτουργίας (που προσδιορίζεται την παρακολούθηση των διακομιστών), τον αριθμό των ενεργών συνδέσεων, τη γεωγραφική θέση, τις δυνατότητες, ή την κυκλοφορία που έχει πρόσφατα ανατεθεί.

2.2.2.4 Χαρακτηριστικά Κατανεμητών Φορτίου

Το υλικό και το λογισμικό των Κατανεμητών Φορτίου μπορεί να έχει μια ποικιλία από ειδικά χαρακτηριστικά. Το θεμελιώδες χαρακτηριστικό της Κατανομής Φόρτου είναι να είναι σε θέση να διανέμει τα εισερχόμενα αιτήματα σε μία σειρά back-end servers του Cluster σύμφωνα με έναν αλγόριθμο προγραμματισμού. Τα περισσότερα από τα ακόλουθα χαρακτηριστικά εξαρτώνται από τον προμηθευτεί:

- Ασύμμετρο Φορτίο (Asymmetric Load): Μια αναλογία φορτίου μπορεί να αντιστοιχιστεί χειροκίνητα για να αναλάβουν κάποιοι backend servers ένα μεγαλύτερο μερίδιο του φόρτου εργασίας απ’ ότι κάποιοι άλλοι . Αυτό μερικές φορές χρησιμοποιείται ως ένας γρήγορος τρόπος για να αναγκάσουμε ορισμένους servers που έχουν μεγαλύτερη χωρητικότητα από άλλους και δεν φέρνουν πάντα τα επιθυμητά αποτελέσματα.

- Ενεργοποίηση Προτεραιότητα (Priority Activation): Όταν ο αριθμός των διαθέσιμων servers πέσει κάτω από ένα ορισμένο αριθμό, ή ο φόρτος που λαμβάνεται είναι πάρα πολύ υψηλός, standby servers μπορούν να συνδεθούν online.

- SSL Offload and Acceleration: Ανάλογα με το φόρτο εργασίας, την επεξεργασία των απαιτήσεων κρυπτογράφησης και ελέγχου ταυτότητας ενός αιτήματος SSL μπορεί να γίνει ένα σημαντικό μέρος στης ζήτηση της CPU του Web Server. Καθώς αυξάνεται η ζήτηση, οι χρήστες θα βλέπουν πιο αργούς χρόνους απόκρισης, όπως το SSL γενικά κατανέμεται μεταξύ των Web servers. Για να καταργηθεί αυτό το αίτημα σε Web servers, ο Κατανεμητής μπορεί να τερματίσει τις συνδέσεις SSL, περνώντας αιτήσεις HTTPS ως αιτήματα HTTP στους Web servers. Αν ο ίδιος ο Κατανεμητής δεν είναι υπερφορτωμένος, αυτό δεν υποβαθμίζει την απόδοση των τελικών χρηστών . Το μειονέκτημα αυτής της προσέγγισης είναι ότι το σύνολο της επεξεργασίας SSL είναι συγκεντρωμένη σε μία μόνο συσκευή (τον Κατανεμητή), η οποία μπορεί να γίνει ένα νέο σημείο συμφόρησης. Μερικές συσκευές Κατανομής Φορτίου περιλαμβάνουν εξειδικευμένο υλικό για την επεξεργασία SSL. Αντί της αναβάθμισης της συντυχίας Κατανομής φορτίου, το οποίο είναι αρκετά ακριβό ειδικό υλικό, είναι φθηνότερο να παραιτηθεί απο SSL Offload και να προσθέσει μερικούς Web servers. Επίσης, ορισμένοι προμηθευτές servers, όπως η Oracle / Sun ενσωματώνουν τώρα ειδικό υλικό επιταχυντή κρυπτογράφησης στους επεξεργαστές τους, όπως ο T2000. Η εταιρία F5 Networks ενσωματώνει μια ειδική κάρτα SSL επιτάχυνσης υλικού σε τοπικό διαχειριστή της κυκλοφορίας LTM (Local Traffic Manager), το οποίο χρησιμοποιείται για την κρυπτογράφηση και αποκρυπτογράφηση της SSL κυκλοφορίας. Ένα σαφές όφελος για SSL εκφόρτωση στον Κατανεμητή είναι ότι δίνει τη δυνατότητα να κάνετε εξισορρόπηση ή μεταγωγή περιεχομένου με βάση τα στοιχεία στην αίτηση HTTPS.

- Distributed Denial of Service (DDoS) επίθεσης και προστασία: Οι Κατανεμητές Φορτίου μπορούν να προσφέρουν χαρακτηριστικά, όπως SYN cookies και καθυστερημένη-δέσμευση (Οι back-end servers δεν βλέπουν τον πελάτη μέχρι να ολοκληρωθεί το TCP handshake «χειραψία») για να μετριάσουν τις επιθέσεις SYN ¨flood¨ και γενικά να αποφορτίσουν από εργασία τους servers σε μια πιο αποδοτική πλατφόρμα.

- HTTP compression: Με την συμπίεση HTTP μειώνει την ποσότητα των δεδομένων που θα μεταφερθούν από HTTP αντικειμένα με τη χρήση συμπίεσης gzip διαθέσιμη σε όλους τους σύγχρονους Web Browsers. Όσο μεγαλύτερη είναι η ανταπόκριση και ποιο μακριά είναι ο πελάτης, τόσο περισσότερο αυτό το χαρακτηριστικό μπορεί να βελτιώσει τους χρόνους απόκρισης. Το δίλημμα είναι ότι αυτό το χαρακτηριστικό βάζει επιπλέον ζήτηση στη CPU του Κατανεμητή Φορτίου και θα μπορούσε να γίνεται από τους Web servers αντ 'αυτού.

- TCP offload: Διαφορετικοί προμηθευτές χρησιμοποιούν διαφορετικούς όρους για το σκοπό αυτό, αλλά η ιδέα είναι ότι συνήθως κάθε αίτηση HTTP από κάθε πελάτη (client) είναι μια διαφορετική σύνδεση TCP. Αυτή η δυνατότητα χρησιμοποιεί HTTP/1.1 για να εδραιώσει πολλαπλές αιτήσεις HTTP από πολλούς πελάτες σε μια ενιαία υποδοχή TCP με τους back-end servers.

- Buffering TCP: Ο κατανεμητής φόρτίου μπορεί να αποθηκεύσει τις αποκρίσεις από το server και να τροφοδοτήσει τα δεδομένα αυτά στους «αργούς» πελάτες , επιτρέποντας στο web server να απελευθερώσει ένα νήμα (thread) για άλλες εργασίες πιο γρήγορα από ό, τι θα ήταν αν έπρεπε να στείλει την αίτηση στο σύνολό της απευθείας στον πελάτη.

- Direct Server Return: Μία επιλογή για ασύμμετρη Κατανομή Φορτίου, όπου το αίτημα και η απάντηση έχουν διαφορετικές διαδρομές στο δίκτυο.

- Health checking: Ο Κατανεμητής περιλαμβάνει servers ελέγχου, για τον έλεγχο του επιπέδου της υγείας και αφαιρεί τους servers που απέτυχαν από την «πισίνα» (πίνακας).

- HTTP caching: Ο Κατανεμητής αποθηκεύει στατικό περιεχόμενο, έτσι ώστε ορισμένα αιτήματα μπορούν να αντιμετωπιστούν χωρίς να επικοινωνήσουν με τους servers.

- Content filtering: Με το φιλτράρισμα περιεχομένου μερικοί Κατανεμητές μπορούν να τροποποιήσουν αυτόνομα την κίνηση την ώρα που βρίσκεται στον δρόμο, δηλαδή δυναμικά.

- HTTP security: Κάποιοι Κατανεμητές μπορεί να κρύψουν σελίδες σφάλματος HTTP, να καταργήσουν επικεφαλίδες ταυτότητας server από HTTP αποκρίσεις και να κρυπτογραφήσουν cookies, έτσι ώστε οι τελικοί χρήστες να μην μπορούν να τα διαχειριστούν.

- Priority queuing: Με τις ουρά προτεραιότητας επίσης γνωστή ως διαμόρφωση ρυθμού, έχουν την ικανότητα να δίνουν διαφορετική προτεραιότητα σε διαφορετικές κυκλοφορίες.

- Content-aware switching: οι περισσότεροι κατανεμητές φορτίου μπορούν να στείλουν αιτήσεις σε διαφορετικούς servers με βάση τη URL που ζητήθηκε, αν υποτεθεί ότι η αίτηση δεν είναι

κρυπτογραφημένη (HTTP) ή, αν είναι κρυπτογραφημένη (μέσω HTTPS) ότι η αίτηση HTTPS τερματίζεται (αποκρυπτογραφηθεί) κατά τη κατανομή φορτίου.

- Client authentication: Η ταυτοποίηση των χρηστών γίνεται από μια ποικιλία πηγών ταυτοποίησης πριν τους επιτραπεί η πρόσβαση σε μια ιστοσελίδα.

- Programmatic traffic manipulation: Με την προγραμματική χειραγώγηση της κυκλοφορίας τουλάχιστον ένας κατανεμητής επιτρέπει τη χρήση μιας γλώσσας προγραμματισμού για να επιτρέψει διάφορες μεθόδους κατανομής φορτίου , μεθόδους δυναμικής χειραγώγησης της κυκλοφορίας, και πολλά άλλα.

- Firewall: Οι απευθείας συνδέσεις με τους backend servers εμποδίζεται, για λόγους ασφάλειας του δικτύου. Firewall είναι ένα σύνολο κανόνων που αποφασίζουν εάν η κίνηση μπορεί να περάσει μέσω μιας διεπαφής ή όχι.

- Instrusion prevenation system: Το σύστημα πρόληψης, προσφέρει ασφάλεια στο επιπέδο εφαρμογής, εκτός από το στρώμα δικτύου/ μεταφοράς που προσφέρει ασφάλεια firewall.

2.2.2.5 Χρήση στις Τηλεπικοινωνίες

Η Κατανομή φορτίου μπορεί να είναι χρήσιμη σε εφαρμογές με περιττούς συνδέσμους επικοινωνίες. Για παράδειγμα, μια εταιρεία μπορεί να έχει πολλαπλές συνδέσεις στο Internet για να εξασφάλιση πρόσβασης στο δίκτυο, αν μία από τις συνδέσεις αποτύχει. Μια διάταξη failover θα σήμαινε ότι μια σύνδεση έχει οριστεί για κανονική χρήση, ενώ ο δεύτερη σύνδεση χρησιμοποιείται μόνο εάν η κύρια σύνδεση αποτύχει.

Χρησιμοποιώντας εξισορρόπηση φορτίου, οι δύο σύνδεσμοι μπορεί να είναι σε χρήση όλη την ώρα. Μια συσκευή ή ένα πρόγραμμα ελέγχει τη διαθεσιμότητα όλων των συνδέσμων και επιλέγει το δρόμο για την αποστολή πακέτων. Η χρήση πολλαπλών δεσμών αυξάνει ταυτόχρονα το διαθέσιμο εύρος ζώνης.

2.2.2.6 Συντομότερη διαδρομή γεφύρωσης (Shortest Path Bridging)

Η IEEE ενέκρινε το IEEE 802.1aq πρότυπο το Μαΐου του 2012,[6] γνωστό και τεκμηριωμένο στα περισσότερα βιβλία, ως Shortest Path Bridging (SPB). Η SPB επιτρέπει σε όλους τους συνδέσμους να είναι ενεργοί για μέσω πολλαπλών διαδρομών ίδιου κόστος, παρέχει ταχύτερους χρόνους σύγκλισης για τη μείωση του χρόνου, και απλοποιεί τη χρήση της εξισορρόπησης φορτίου σε τοπολογίες δικτύου πλέγματος (μερικώς συνδεδεμένα ή και πλήρως συνδεδεμένη), επιτρέποντας να φορτώνει μερίδιο της κυκλοφορίας σε όλα μονοπάτια ενός δίκτυο.[7][8] Η SPB έχει σχεδιαστεί για να εξαλείψει ουσιαστικά το ανθρώπινο λάθος κατά τη διαμόρφωση και διατηρεί την plug-and-play φύση του οπου ιδρύθηκε το Ethernet ως το de facto πρωτόκολλο στο Επίπεδο 2.[9]

2.2.2.7 Δρομολόγηση (Routing)

Πολλές εταιρείες τηλεπικοινωνιών έχουν πολλαπλές διαδρομές μέσω των δικτύων τους ή μέσα από εξωτερικά δίκτυα. Χρησιμοποιούν εξελιγμένους κατανεμητές φορτίου για την μετατόπιση της κυκλοφορίας από τη μία διαδρομή στην άλλη για να αποφευχθεί η συμφόρηση του δικτύου σε κάποια συγκεκριμένη σύνδεση και μερικές φορές για να ελαχιστοποιηθεί το κόστος της διέλευσης από τα εξωτερικά δίκτυα ή τη βελτίωση της αξιοπιστίας του δικτύου.

Ένας άλλος τρόπος χρήσης της εξισορρόπησης φορτίου είναι σε παρακολούθηση των δραστηριοτήτων του δικτύου (monitoring). Ο κατανεμητής φορτίου μπορεί να χρησιμοποιηθεί για να χωρίσει τεράστιες ροές δεδομένων σε πολλές επιμέρους ροές δεδομένων καθώς και να χρησιμοποιηθούν διάφοροι αναλυτές δικτύου, καθένας διαβάζοντας ένα μέρος των αρχικών δεδομένων. Αυτό είναι πολύ χρήσιμο για την παρακολούθηση γρήγορων δικτύων όπως το 10GbE ή STM64, όπου η σύνθετη επεξεργασία των δεδομένων μπορεί να μην είναι δυνατή με την ταχύτητα που έχει το μέσο μεταφοράς.

2.2.3 Συστοιχίες υψηλής απόδοσης (High Performance Computing Clusters)

Πολλοί οργανισμοί έχουν μεγάλες ποσότητες δεδομένων που έχουν συλλεχθεί και αποθηκευτεί σε μεγάλα σύνολα δεδομένων που πρέπει να υποβληθούν σε επεξεργασία και ανάλυση για την παροχή επιχειρηματικών πληροφοριών, τη βελτίωση των προϊόντων και των υπηρεσιών για τους πελάτες, ή για να ικανοποιήσουν άλλες εσωτερικές απατήσεις επεξεργασίας των δεδομένων. Για παράδειγμα, εταιρείες Διαδικτύου πρέπει να επεξεργάζονται τα δεδομένα που συλλέγονται από το διαδίκτυο, καθώς και τα αρχεία καταγραφής και άλλες πληροφορίες που προέρχονται από τις υπηρεσίες Web. Παράλληλες σχεσιακές βάσης

δεδομένων της τεχνολογίας δεν έχουν αποδειχθεί ότι είναι οικονομικά αποδοτικές ή να παρέχουν την υψηλή απόδοση που απαιτείται για την ανάλυση τεράστιων ποσοτήτων δεδομένων σε εύθετο χρόνο. Ως αποτέλεσμα πολλοί οργανισμοί ανέπτυξαν την τεχνολογία για την αξιοποιήσει μεγάλων Cluster αποτελούμενα από servers για να παρέχουν υψηλής απόδοσης δυνατοτήτων αξιοποίησης της υπολογιστικής ισχύς και για την επεξεργασία και την ανάλυση μεγάλων συνόλων δεδομένων. Clusters μπορεί να αποτελούνται από εκατοντάδες ή ακόμα και χιλιάδες εμπορικά μηχανήματα-υπολογιστές όπου συνδέονται με τη χρήση των δικτύων υψηλού εύρους ζώνης (High Bandwith Networks). Παραδείγματα αυτού του τύπου της τεχνολογίας Cluster περιλαμβάνονται στο MapReduce της Google, Apache Hadoop, Aster Data Systems, Sector / Sphere και LexisNexis HPCC πλατφόρμα**.**

2.2.3.1 High Performance Computing (HPC)

Ο όρος High Performance Computing (HPC) αναφέρεται σε υπολογιστικά περιβάλλοντα που χρησιμοποιούν οι υπερυπολογιστές και συστοιχίες Cluster για την αντιμετώπιση πολύπλοκων υπολογιστικών απαιτήσεων, υποστήριξη εφαρμογών με πολύ μεγάλες χρονικές απαιτήσεις επεξεργασίας, ή να απαιτούν την επεξεργασία σημαντικών ποσοτήτων δεδομένων. Οι υπερυπολογιστές γενικά σχετίζονται με την επιστημονική έρευνα και τον εντατικό υπολογισμό δύσκολων προβλημάτων, αλλά όλο και περισσότερο η τεχνολογία των υπερυπολογιστών προορίζετε και για εντατικούς υπολογισμούς αλλά και για υπολογισμό εντατικών δεδομένων. Μια νέα τάση στο σχεδιασμό υπερυπολογιστών για High Performance Computing είναι χρησιμοποιώντας Clusters με ανεξάρτητους επεξεργαστές σε παράλληλη σύνδεση. Πολλά προβλήματα υπολογισμών είναι κατάλληλα για παράλληλη επεξεργασία, συχνά τα προβλήματα μπορούν να διαιρεθούν κατά τέτοιο τρόπο ώστε κάθε ανεξάρτητος κόμβος επεξεργασίας μπορεί να επεξεργαστεί ένα τμήμα του προβλήματος παράλληλα απλά διαιρώντας τα δεδομένα που πρόκειται να επεξεργαστούν και στη συνέχεια συνδυάζοντας τα τελικά αποτελέσματα επεξεργασίας από κάθε τμήμα. Αυτό το είδος της παραλληλισμού συχνά αναφέρεται ως Παραλληλισμός-δεδομένων (Data-Parallelism).

Οι εφαρμογές Παραλληλισμού-Δεδομένων είναι μια πιθανή λύση για τις απαιτήσεις επεξεργασίας δεδομένων σε μεγάλη κλίμακα, όπως είναι η petabyte. Ο παραλληλισμός-δεδομένων μπορεί να οριστεί ως ένας υπολογισμός που εφαρμόζεται μεμονωμένα σε κάθε στοιχείο δεδομένων από ένα σύνολο δεδομένων το οποίο επιτρέπει τον βαθμό παραλληλισμού για να κλιμακωθεί κατάλληλα με τον όγκο των δεδομένων. Ο πιο σημαντικός λόγος για την ανάπτυξη εφαρμογών Παραλληλισμού-Δεδομένων είναι η δυνατότητα για επέκταση των

επιδόσεων των συστημάτων High Performance Computing (HPC), όπου μπορεί να οδηγήσει σε μεγάλο βαθμό στην βελτίωση των επιδόσεών τους.

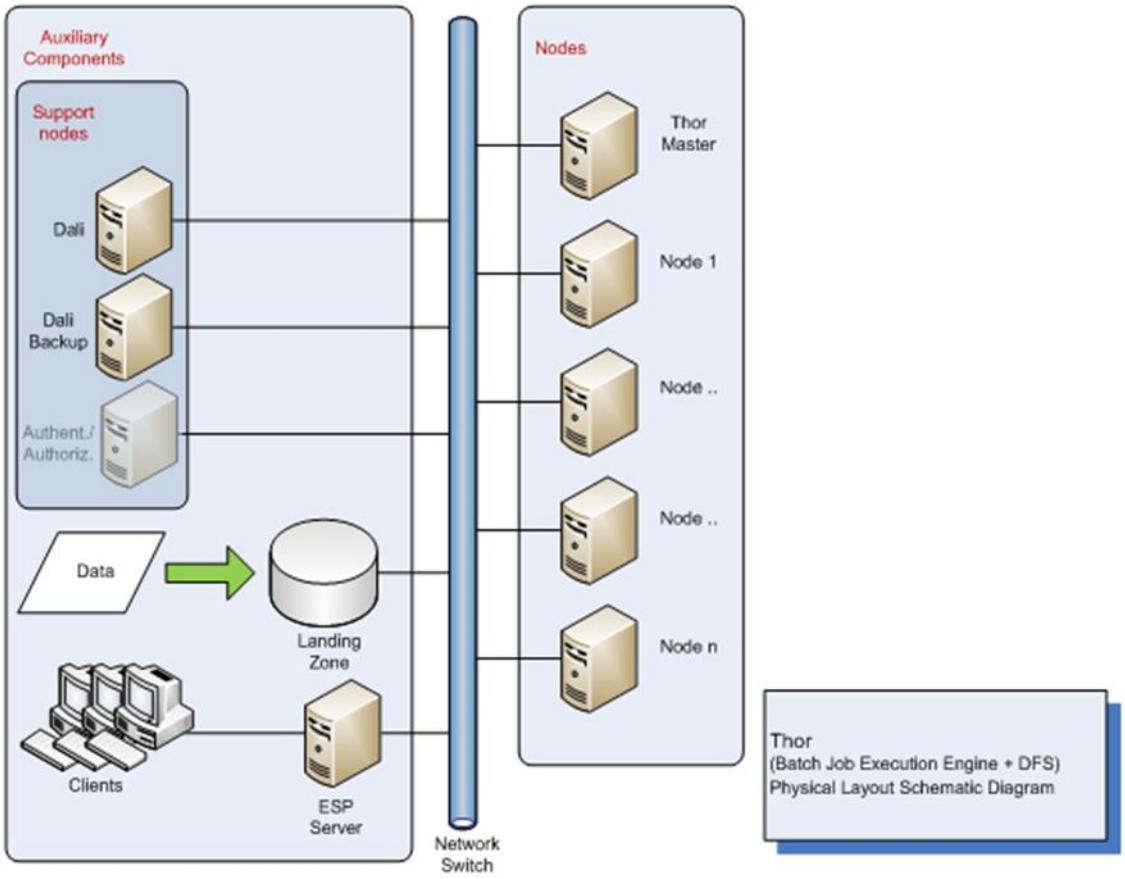
2.2.3.2 Αρχιτεκτονική συστήματος

Η αρχιτεκτονική του συστήματος HPCC περιλαμβάνει δύο διακριτά περιβάλλοντα επεξεργασίας Cluster, καθένα από τα οποία μπορεί να βελτιστοποιηθεί ανεξάρτητα για παράλληλου σκοπού επεξεργασίας των δεδομένων του. Η πρώτη από αυτές τις πλατφόρμες ονομάζεται «διυλιστήριο δεδομένων» (Data refinery) του οποίου γενικός σκοπός είναι η γενική επεξεργασία τεράστιων όγκων ανεπεξέργαστα δεδομένα οποιουδήποτε τύπου για οποιοδήποτε σκοπό, αλλά χρησιμοποιείται συνήθως για τον καθαρισμό και την υγεία των δεδομένων, να εξάγει, να μετατρέπει, φόρτοση επεξεργασίας ακατέργαστων δεδομένων, καταγραφή σύνδεση και μεγάλης κλίμακας ad-hoc ανάλυσης στοιβάδων δεδομένων, και τη δημιουργία, εισάγονται στοιχεία και δείκτες για την υποστήριξη υψηλής απόδοσης δομημένων ερωτημάτων και εφαρμογές αποθήκευσης δεδομένων. Το Data Refinery επίσης αναφέρεται ως Thor, αναφορά στη μυθική μορφή του Νορβηγού θεού, του κεραυνού με το μεγάλο σφυρί να είναι συμβολικό για τη σύνθλιψη μεγάλων ποσοτήτων ακατέργαστων δεδομένων σε χρήσιμες πληροφορίες. Ένα Thor Cluster είναι παρόμοιο σε λειτουργία, περιβάλλον εκτέλεσης, σύστημα αρχείων και τις δυνατότητες της Google και Hadoop MapReduce πλατφόρμας .

Το Σχήμα 2.5 δείχνει μια απεικόνιση ενός φυσικού Thor cluster επεξεργασίας που λειτουργεί ως μια δέσμη μηχανή εκτέλεσης εργασίας για υπολογισμό εφαρμογών με μεγάλης κλίμακας δεδομένα.

Εκτός από τους Thor master και slave κόμβους, χρειάζονται πρόσθετα βοηθητικά κοινά στοιχεία για να λειτουργήσει ένα ολοκληρωμένο περιβάλλον επεξεργασίας HPCC. Η δεύτερη από τις παράλληλες πλατφόρμες επεξεργασίας δεδομένων ονομάζεται Roxie και λειτουργεί ως μηχανή ταχείας παράδοσης δεδομένων. Αυτή η πλατφόρμα έχει σχεδιαστεί ως μία online υψηλής απόδοσης δομημένη αναζήτηση και πλατφόρμα ανάλυσης ή αποθήκευσης δεδομένων παρέχοντας παράλληλα τις απαιτήσεις πρόσβασης επεξεργασίας στα δεδομένα των online εφαρμογών μέσω των υπηρεσιών Web υπηρεσιών που υποστηρίζουν χιλιάδες ταυτόχρονα αιτήματα και χρήστες με χρόνους απόκρισης κλάσματος του δευτερολέπτου. Το Roxie χρησιμοποιεί ένα κατανεμημένο σύστημα αρχείων με ευρετήριο, για την παροχή παράλληλης επεξεργασίας των αιτημάτων χρησιμοποιώντας ένα βελτιστοποιημένο περιβάλλον εκτέλεσης και αρχείων για υψηλής απόδοσης online επεξεργασία. Ένα Roxie cluster είναι παρόμοιο σε κάθε λειτουργία και δυνατότητες για Hadoop με HBase και πρόσθετες δυνατότητες Hive και

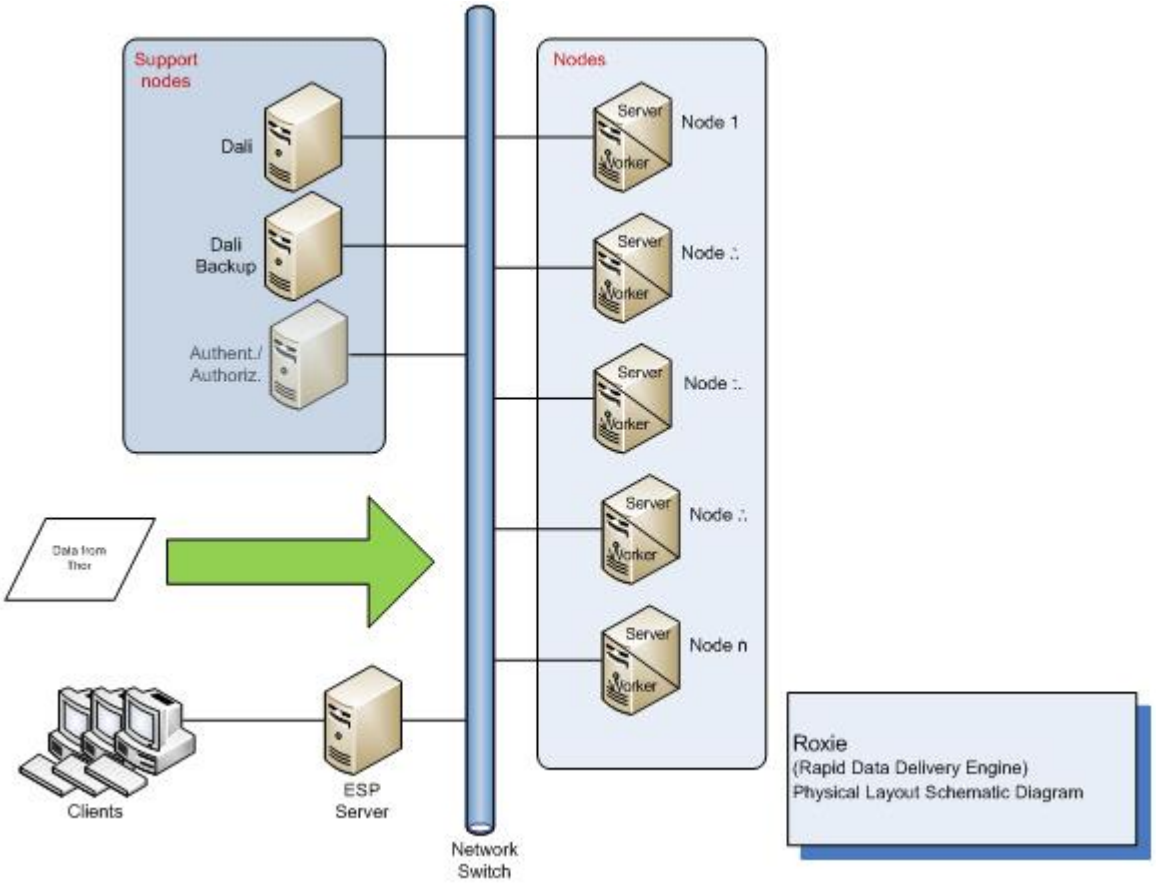
παρέχει σχεδόν σε πραγματικό χρόνο προβλέψεις αιτημάτων. Και οι Thor και οι Roxie clusters χρησιμοποιούν τη γλώσσα προγραμματισμού ECL για την υλοποίηση εφαρμογών, την αύξηση της συνέχειας και της παραγωγικότητας του προγραμματιστή.



Σχήμα: 2.5 Thor Processing Cluster

Το Σχήμα 2.6 δείχνει μια απεικόνιση ενός φυσικού Roxie cluster επεξεργασίας που λειτουργεί ως μία online μηχανή εκτέλεσης αιτημάτων για τα αιτήματα υψηλών απαιτήσεων και των εφαρμογών αποθήκευσης δεδομένων. Ένα Roxie cluster περιλαμβάνει πολλαπλούς κόμβους με διακομιστή (Server) και διακομιστές/εργάτες (server/workers) γα επεξεργασία των αιτημάτων. Ένα πρόσθετο βοηθητικό συστατικό που ονομάζεται διακομιστής (Serveer) ESP το οποίο παρέχει διεπαφές για εξωτερική πρόσβαση του πελάτη στο cluster και επιπλέον κοινά στοιχεία τα οποία είναι κοινά με ένα Thor cluster σε ένα περιβάλλον HPCC. Παρά το γεγονός ότι ένα Thor cluster επεξεργασίας μπορεί να υλοποιηθεί και να χρησιμοποιηθεί χωρίς ένα Roxie cluster, ένα περιβάλλον HPCC που περιλαμβάνει ένα Roxie cluster θα πρέπει επίσης να περιλαμβάνει ένα Thor cluster. Το Thor cluster χρησιμοποιείται για την κατασκευή των δεικτών κατανεμημένων αρχείων που χρησιμοποιούνται από το Roxie cluster και για την

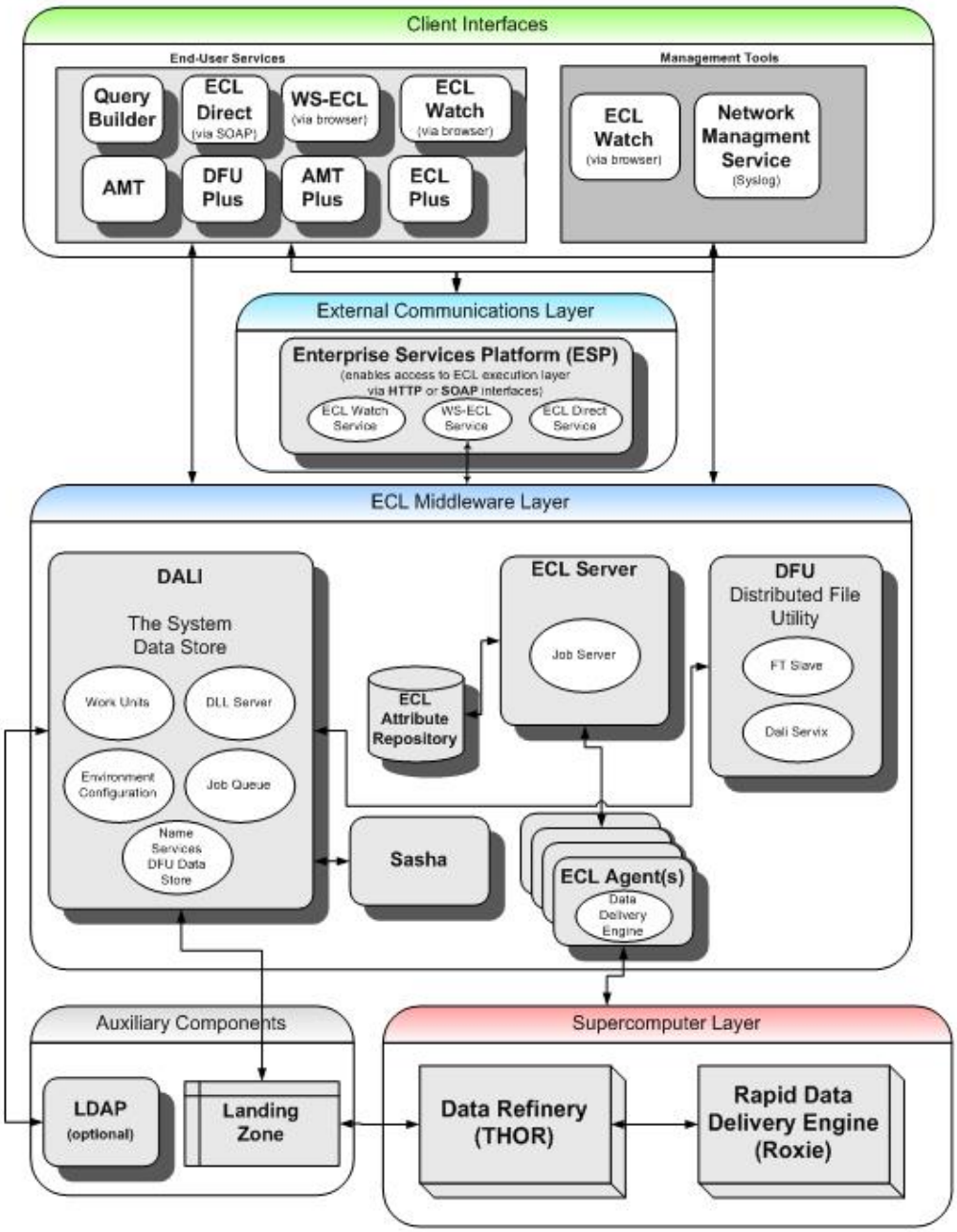
ανάπτυξη online αιτημάτων τα οποία θα πρέπει να συνδεθούν με τους δείκτες των αρχείων στο Roxie cluster.



Σχήμα:2.6 Roxie processing cluster

2.2.3.3 Αρχιτεκτονική Λογισμικού (Software Architecture)

Η αρχιτεκτονική λογισμικού HPCC ενσωματώνει στα Thor και Roxie clusters καθώς και κοινά τους συστατικά middleware, ένα εξωτερικό στρώμα επικοινωνίας, διεπαφές πελάτη που παρέχουν τόσο υπηρεσίες τελικών χρηστών και τα εργαλεία διαχείρισης του συστήματος, καθώς και βοηθητικά στοιχεία για την υποστήριξη της παρακολούθησης και για να διευκολύνεται η φόρτωση και η αποθήκευση των αρχείων δεδομένων από εξωτερικές πηγές. Ένα περιβάλλον HPCC μπορεί να περιλαμβάνει μόνο Thor clusters, ή και τα δύο, Thor και Roxie clusters. Η συνολική αρχιτεκτονική του λογισμικού HPCC δείχνεται στο Σχήμα 2.7.



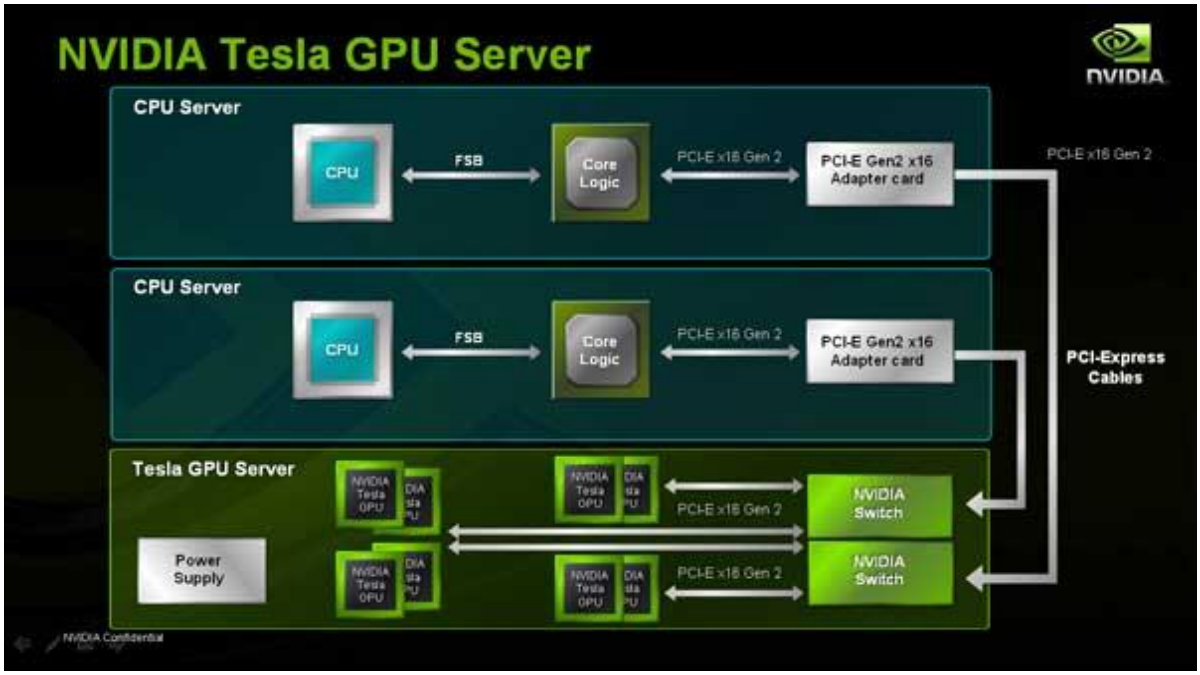
Σχήμα: 2.7 HPCC Software architecture

2.3 Σχεδιασμός και Διαμόρφωση

Ένα από τα ζητήματα στο σχεδιασμό ενός cluster είναι πόσο στενά συνδεμένοι μπορεί να είναι οι επιμέρους κόμβοι. Για παράδειγμα, μία απλή εργασία υπολογιστή μπορεί να απαιτεί συχνή επικοινωνία μεταξύ των κόμβων: αυτό σημαίνει ότι το Cluster μοιράζεται ένα αποκλειστικό δίκτυο, το οποίο είναι τοπικό και πιθανώς έχει ομοιογενείς κόμβους. Στο άλλο άκρο είναι μια εργασία όπου ο υπολογιστής χρησιμοποιεί έναν ή λίγους κόμβους και χρειάζεται

λίγη ή καθόλου επικοινωνία μεταξύ των κόμβων. Αυτή η προσέγγιση πλησιάζει την τεχνολογία υπολογιστικού πλέγματος (Grid computing).

Σε ένα σύστημα Beowulf, τα προγράμματα εφαρμογής ποτέ δεν βλέπουν τους υπολογιστές κόμβους (που ονομάζεται επίσης και Slave computers), αλλά αλληλοεπιδρούν μόνο με τον "Master" ο οποίος είναι ένας συγκεκριμένος υπολογιστής όπου αναλαμβάνει τον προγραμματισμό και της διαχείριση των Slave computers.[10] Σε μια τυπική εφαρμογή ο Master έχει δύο διεπαφές δικτύου, μια που επικοινωνεί με το ιδιωτικό δίκτυο Beowulf για τους Slaves και από την άλλη για το δίκτυο γενικού σκοπού του οργανισμού ή το Διαδίκτυο.[10] Οι Slave computers έχουν συνήθως τη δική τους εκδοχή του ίδιου του λειτουργικού συστήματος, καθώς και την τοπική μνήμη και σκληρό δίσκο. Ωστόσο, το ιδιωτικό δίκτυο “slave” μπορεί επίσης να έχει έναν μεγάλο και κοινόχρηστο διακομιστή αρχείων (File Server) που αποθηκεύει παγκόσμια σταθερά δεδομένα, τα οποία είναι προσβάσιμα από τους Slaves computers, όταν τα ζητούν.[10] Αντίθετα, ο ειδικού σκοπού DEGIMA Cluster των 144 κόμβων, είναι συντονισμένοι για να «τρέχουν» αστροφυσικές προσομοιώσεις N-σωμάτων χρησιμοποιώντας το Multiple-Walk parallel treecode, αντί για γενικής χρήσης επιστημονικούς υπολογισμούς.[11] Λόγω της αυξανόμενης υπολογιστικής ισχύος της κάθε γενιάς κονσόλας παιχνιδιών, μια νέα χρήση έχει προκύψει όπου αναδιαμορφώνει τις κονσόλες σε Cluster υπολογιστές υψηλής απόδοσης (HPC) . [12]



Σχήμα: 2.8 Nvidia Tesla Personal Supercomputer workstation

Μερικά παραδείγματα με κονσόλες clusters είναι το Sony PlayStation Cluster και το Microsoft Xbox cluster. Ένα άλλο παράδειγμα είναι το Nvidia Tesla Personal Supercomputer workstation, το οποία χρησιμοποιεί πολλαπλούς επεξεργαστές από κάρτες γραφικών της Nvidia. [13]

2.3.1 PlayStation 3 Cluster

Η σημαντική υπολογιστική ικανότητα των μικροεπεξεργαστών του PlayStation 3 κέντρισε το ενδιαφέρον σε πολλούς χρήστες. Χρησιμοποιώντας τα PS3 σε δίκτυο για διάφορες εργασίες που απαιτούν έναν προσιτό Υψηλών επιδόσεων υπολογιστή(HPC). Η NCSA είχε ήδη κατασκευάσει ένα Cluster που βασίζεται στο PlayStation 2. [14] Η Terra Soft Solutions έχει μια έκδοση του Yellow Dog Linux για το PlayStation3 [15] και πουλάει PS3 με το Linux προεγκατεστημένα, σε μεμονωμένες μονάδες, σε 8 και 32 κόμβων Cluster.[16] Επιπλέον, η RapidMind προωθεί το stream programming package για το PS3. [17][18]

Στις 3 Ιανουαρίου 2007, ο Δρ Frank Mueller, Αναπληρωτής Καθηγητής της Επιστήμης Υπολογιστών στο NCSU, Κατασκεύασε ένα Cluster με 8 PS3. Ο Mueller σχολίασε ότι τα 256 MB RAM συστήματος είναι ένας περιορισμός για τη συγκεκριμένη κατασκευή, και εξέτασαν το ενδεχόμενο να προσπαθήσουν να αναβαθμισουν με περισσότερη μνήμη RAM. Το λογισμικό περιλαμβάνει: Fedora Core 5 Linux ppc64, MPICH2, OpenMP v2.5, GNU Compiler Collection και CellSDK 1.1.[19][20][21] Το καλοκαίρι του 2007, ο Gaurav Khanna, καθηγητής στο Τμήμα Φυσικής του Πανεπιστημίου της Μασαχουσέτης Dartmouth κατασκεύασε ένα ανεξάρτητο cluster message-passing βασιζόμενο σε 8 PS3 και τρέχει Fedora Linux.[22] Αυτό το cluster δημιουργήθηκε με την υποστήριξη της Sony Computer Entertainment και ήταν το πρώτο τέτοιου είδους Cluster που δημιουργείται για την δημοσίευση επιστημονικών αποτελεσμάτων. Αποκαλούμενο ως το "PS3 Gravity Grid", αυτό το PS3 cluster εκτελεί αστροφυσικές προσομοιώσεις για μεγάλες «μαύρες τρύπες» οι οποίες συλλαμβάνουν μικρότερα συμπαγή αντικείμενα. Ο Khanna ισχυρίζεται ότι οι επιδόσεις του Cluster υπερβαίνει τις δυνατότητες 100 + Intel Xeon επεξεργαστών όπου βασίζεται παραδοσιακά το cluster Linux σε προσομοιώσεις του . Το PS3 Gravity Grid συγκεντρώθηκαν σημαντική προσοχή των μέσων ενημέρωσης στα μέσα του 2007,[23][24] το 2008,[25][26] το 2009[27][28][29] και το 2010.[30][31] Ο Khanna δημιούργησε επίσης μία DIY ιστοσελίδα για το πώς να οικοδομήσουν τέτοια cluster, ώστε να είναι προσβάσιμα στο ευρύ κοινό. Τον Νοέμβριο του 2010 το Εργαστήριο Air Force Research (AFRL) δημιούργησε ένα ισχυρό υπερυπολογιστή συνδέοντας μεταξύ τους 1,760 Sony PS3 που περιλαμβάνουν 168 ξεχωριστές μονάδες

επεξεργασίας γραφικών και για το συντονισμό 84 servers σε παράλληλη συστοιχία. Το cluster είναι ικανό να εκτελεί 500 τρισεκατομμύρια πράξεις κινητής υποδιαστολής ανά δευτερόλεπτο (500 TFLOPS).[32] Έτσι χτίστηκε το Condor Cluster ήταν το 33ο μεγαλύτερος υπερυπολογιστής στον κόσμο και θα μπορούσε να χρησιμοποιηθεί για την ανάλυση εικόνων υψηλής ευκρίνειας από δορυφόρους.[33]

2.3.1.1 Χρήση σε Ιατρικές έρευνες

Στις 22 Μαρτίου του 2007, το SCE και το Πανεπιστήμιο του Stanford επέκτεινε το Folding @ home project στο PS3.[34] Μαζί με χιλιάδες υπολογιστές που έχουν ήδη ενταχθεί μέσω του Διαδικτύου, οι κάτοχοι PS3 μπορούν να δανείσουν την υπολογιστική ισχύ των συστημάτων τους (κονσόλες) στη μελέτη της ακατάλληλης αναδίπλωσης της πρωτεΐνης και των συναφών ασθενειών, όπως Alzheimer, Parkinson, Huntington, κυστική ίνωση, και διάφορες μορφές καρκίνου. Το λογισμικό είχε συμπεριληφθεί ως μέρος της ενημέρωσης 1.6 firmware (22 Μάρτίου του 2007), και μπορεί να ρυθμιστεί ώστε να λειτουργεί χειροκίνητα ή αυτόματα, όταν το PS3 είναι σε αδράνεια μέσω του Cross Media Bar. Η επεξεργασία πληροφοριών στη συνέχεια στέλνεται πίσω στους κεντρικούς servers του έργου μέσω του Διαδικτύου. Η Επεξεργαστική ισχύ από τους χρήστες PS3 συμβάλλει σε μεγάλο βαθμό στο έργο Folding @ home. [35] Από το Μάρτιο του 2011, περισσότεροι από ένα εκατομμύριο ιδιοκτήτες PS3 έχουν επιτρέψει το Folding @ Home λογισμικό να τρέξει στα συστήματά τους, με πάνω από 27.000 να δραστηριοποιούνται σήμερα, για ένα σύνολο 8,1 petaflops. Συγκριτικά ο ισχυρότερος υπερυπολογιστής του κόσμου, μέχρι τον Νοέμβριο του 2010, ήταν ο Tianhe-IA που έχει μέγιστη απόδοση 2,56 petaflops, ή 2566 teraflops.[36] Η τελευταία έκθεση αναφέρει ότι Folding @ Home έχει περάσει το όριο του 5 petaflop, εκ των οποίων τα 767 teraflops παρέχονται από χρήστες του PlayStation 3 .

Το εργαστήριο Υπολογιστικής Βιοχημείας και Βιοφυσικής στη Βαρκελώνη έχει ξεκινήσει ένα ερευνητικό πρόγραμμα κατανεμημένων υπολογιστών το οποίο ονομάζεται PS3GRID. Το συγκεκριμένο έργο αναμένεται να «τρέχει» δεκαέξι φορές πιο γρήγορα από ό, τι ένα ισοδύναμο πρόγραμμα σε ένα τυποποιημένο PC. Όπως τα περισσότερα έργα κατανεμημένης επεξεργασίας, είναι σχεδιασμένο να λειτουργεί μόνο όταν ο υπολογιστής είναι αδρανής.

Το eHiTS Lightning είναι η πρώτη εικονική απεικόνιση και λογισμικό για το PS3.[37] Διατέθηκε από τη SimBioSys.[38] Όπως αναφέρθηκε από την Bio-IT World τον Ιούλιο του 2008.[39] Αυτή η εφαρμογή τρέχει μέχρι και 30x γρηγορότερα σε ένα PS3 από ό, τι σε ένα

κανονικό PC με επεξεργαστή ενός πυρήνα. Τρέχει και σε PS3 Cluster, επιτυγχάνοντας τον έλεγχο των τεράστιων βιβλιοθηκών χημικής ένωσης σε λίγες ώρες ή ημέρες και όχι εβδομάδες.

2.3.1.2 Η Ματαίωση του PlayStation 3 Cluster

Στις 28 Μαρτίου, 2010, η Sony ανακοίνωσε ότι θα πρέπει να απενεργοποιήσει τη δυνατότητα να τρέξει άλλο λειτουργικό σύστημα με την ενημερωμένη έκδοση V3.21, λόγω ανησυχιών για την ασφάλεια σχετικά με OtherOs.[40] Αυτή η ενημερωμένη έκδοση δεν επηρέασε τα υπάρχοντα cluster υπερυπολογιστών, αυτό οφείλεται στο γεγονός ότι δεν συνδέονται με PSN και δεν τα αναγκάζει να προβούν σε ενημερώσεις (updates) . Ωστόσο, αυτό κάνει την αντικατάσταση των ατομικών playstation που συνθέτουν τα clusters πολύ δύσκολο, αν όχι αδύνατο, δεδομένου ότι τυχόν νεότερα μοντέλα με τη V3.21 ή νεότερη έκδοση δεν υποστηρίζουν την εγκατάσταση του Linux άμεσα.[41] Αυτό προκάλεσε το τέλος κοινή χρήση του PS3 για clustering, αν και υπάρχουν προγράμματα όπως το "The Condor» που εξακολουθεί να δημιουργείται με παλαιότερες μονάδες PS3, και έχουν έρθει online μετά την 1, Απριλίου 2010. [42]

2.4 Διαμοιρασμός δεδομένων και επικοινωνία

Δεδομένου ότι τα computer cluster εμφανίστηκαν κατά τη διάρκεια της δεκαετίας του 1980, οπότε ήταν υπερυπολογιστές. Ένα από τα στοιχεία που διέκρινε τις τρεις τάξεις εκείνη την εποχή ήταν ότι οι πρώτοι υπερυπολογιστές στηρίζονταν στον διαμοιρασμό της μνήμης. Σήμερα τα cluster δεν χρησιμοποιούν συνήθως φυσική κοινόχρηστη μνήμη, ενώ πολλές παρόμοιες αρχιτεκτονικές υπερυπολογιστή έχουν επίσης εγκαταλειφθεί.

Ωστόσο, η χρήση αρχείων συστημάτων cluster (Clustered file system) είναι απαραίτητη στα σύγχρονα computer cluster. Παραδείγματα περιλαμβάνουν την IBM με το General Parallel File System, το Cluster Shared Volumes της Microsoft και το Oracle Cluster File System.

2.5 Message passing and communication

Δύο ευρέως χρησιμοποιούμενες προσεγγίσεις για την επικοινωνία μεταξύ των κόμβων του cluster είναι η MPI, το Message Passing Interface και η PVM, η Parallel Virtual Machine.[43]

Το PVM αναπτύχθηκε στο Εθνικό Εργαστήριο Oak Ridge γύρω στο 1989, πριν το MPI να είναι διαθέσιμο. Το PVM πρέπει να εγκατασταθεί άμεσα σε κάθε κόμβο του cluster και παρέχει μια σειρά από βιβλιοθήκες λογισμικού που «βαφτίζουν» τον κόμβο ως «εικονική παράλληλη μηχανή". Το PVM παρέχει ένα περιβάλλον χρόνου εκτέλεσης για το message passing, το έργο και τη διαχείριση των πόρων και την κοινοποίηση του σφάλματος. Το PVM μπορεί να χρησιμοποιηθεί από τα προγράμματα του χρήστη γραμμένα σε C, C + + ή Fortran, κλπ.[43][45]

Το MPI εμφανίσθηκε στις αρχές της δεκαετίας του 1990 από τις συζητήσεις ανάμεσα σε 40 οργανώσεις. Η αρχική προσπάθεια υποστηρίχθηκε από την ARPA και το Εθνικό Ίδρυμα Επιστημών. Αντί να ξεκινούν από την αρχή, ο σχεδιασμός του MPI βασίστηκε σε διάφορα χαρακτηριστικά που είναι διαθέσιμα σε εμπορικά συστήματα της εποχής. Οι προδιαγραφές του MPI στη συνέχεια οδήγησε σε συγκεκριμένες υλοποιήσεις. Οι υλοποιήσεις του MPI συνήθως χρησιμοποιούν το πρωτόκολλο TCP / IP και συνδέσεις υποδοχής.[46] Το MPI είναι πλέον ένα ευρέως διαθέσιμο μοντέλο επικοινωνίας που επιτρέπει παράλληλα προγράμματα να γραφτούν σε γλώσσες όπως η C, Fortran, Python, κλπ.[47] Έτσι, σε αντίθεση με το PVM που παρέχει μια συγκεκριμένη εφαρμογή, το MPI είναι μια προδιαγραφή που έχει εφαρμοστεί σε συστήματα όπως MPICH και Open MPI.[48][49]

2.6 Διαχείριση Cluster

Μία από τις προκλήσεις στη χρήση ενός computer cluster είναι το κόστος της διαχείρισης, όπου αυτό μπορεί μερικές φορές να είναι τόσο υψηλό όσο το κόστος της διαχείρισης των N ανεξάρτητων μηχανημάτων, αν το σύμπλεγμα έχει N κόμβους.[50] Σε ορισμένες περιπτώσεις, αυτό αποτελεί ένα πλεονέκτημα σε κοινές αρχιτεκτονικές μνήμης με χαμηλότερο κόστος διαχείρισης.[50] Αυτό έχει επίσης κάνη δημοφιλή τις εικονικές μηχανές, λόγω της ευκολίας της διαχείρισης .[50]

2.7 Προγραμματισμός διεργασιών

Όταν ένα μεγάλο cluster πολλαπλών χρηστών πρέπει να έχει πρόσβαση σε πολύ μεγάλες ποσότητες δεδομένων, ο προγραμματισμός διεργασιών γίνεται μια πρόκληση. Σε ένα ετερογενές cluster CPU-GPU, το οποίο έχει ένα σύνθετο περιβάλλον εφαρμογής, η εκτέλεση κάθε εργασίας εξαρτάται από τα χαρακτηριστικά των υποκείμενων του cluster, εργασίες χαρτογράφησης των πυρήνων CPU και GPU είναι συσκευές που παρέχουν σημαντικές προκλήσεις.[51] Πρόκειται για έναν τομέα της συνεχούς έρευνας όπου αλγόριθμοι

συνδυάζονται για να επεκταθεί το MapReduce και το Hadoop και έχουν προταθεί και μελετηθεί.[51]

2.8 Διαχείριση κόμβων που απέτυχαν

Όταν ένας κόμβος σε ένα cluster αποτύχει, μπορούν να χρησιμοποιηθούν στρατηγικές όπως «περίφραξη» για να κρατήσει το υπόλοιπο του συστήματος σε λειτουργία.[52][53] Η περίφραξη είναι η διαδικασία απομόνωσης ενός κόμβου ή προστασία κοινών πόρων, όταν ένας κόμβος φαίνεται να έχει δυσλειτουργία. Υπάρχουν δύο κατηγορίες μεθόδων περίφραξης. Η πρώτη είναι η απενεργοποιήσει του ίδιου του κόμβου από μόνος του, και οι άλλοι κόμβοι απορρίπτουν τη πρόσβαση σε πόρους, όπως κοινόχρηστους δίσκους.[52] Η μέθοδος STONITH σημαίνει "Shoot The Other Node In The Head", που σημαίνει ότι ο ύποπτος κόμβος είναι απενεργοποιημένος ή είναι σβηστός.[52] Για παράδειγμα, η περίφραξη ενέργειας χρησιμοποιεί έναν ελεγκτή λειτουργίας για να απενεργοποιήσει τον κόμβο που έχει πρόβλημα.

2.9 Ανάπτυξη Λογισμικού και Διαχείριση

2.9.1 Παράλληλος προγραμματισμός

Τα Cluster εξισορρόπησης φορτίου (Load Balancing Clusters), όπως web servers χρησιμοποιούν αρχιτεκτονικές cluster για να υποστηρίξουν ένα μεγάλο αριθμό χρηστών και τυπικά κάθε αίτημα του χρήστη δρομολογείται σε ένα συγκεκριμένο κόμβο, επιτυγχάνοντας παραλληλισμό της εργασία χωρίς συνεργασία των κόμβων, δεδομένου ότι ο κύριος στόχος του συστήματος είναι η παροχή ταχείας πρόσβασης στον χρήστη σε κοινά δεδομένα. Ωστόσο, τα computer cluster που εκτελούν περίπλοκους υπολογισμούς για ένα μικρό αριθμό χρηστών πρέπει να επωφεληθούν από τις δυνατότητες παράλληλης επεξεργασίας του cluster και να διαμοιράσουν σε τμήματα «τον ίδιο υπολογισμό» μεταξύ πολλών κόμβων.[53]

Η αυτόματη παραλληλοποίηση (Automatic Paralillezation) των προγραμμάτων εξακολουθεί να παραμένει μια τεχνική πρόκληση, αλλά μοντέλα παράλληλου προγραμματισμού μπορούν να χρησιμοποιηθούν για να υλοποιήσουν ένα υψηλότερο βαθμό παραλληλισμού με την ταυτόχρονη εκτέλεση των ξεχωριστών τμημάτων ενός προγράμματος σε διαφορετικούς επεξεργαστές-κόμβους.[53][54]

2.9.2 Αποσφαλμάτωση και παρακολούθηση

Η ανάπτυξη και η αποσφλμάτωση(debugging) των παράλληλων προγραμμάτων σε ένα cluster απαιτεί θεμελιακή παράλληλη γλώσσα, καθώς και τα κατάλληλα εργαλεία, όπως αυτά που συζητήθηκαν στο High Performance Debugging Forum (hpdf), τα οποία οδήγησαν στη συγγραφή του HPD.[55][56] Εργαλεία όπως είναι το TotalView όπου στη συνέχεια αναπτύχθηκε για τον εντοπισμό σφαλμάτων παράλληλων υλοποιήσεων σε cluster computers που χρησιμοποιούν MPI ή PVM ή για τη μετάδοση μηνυμάτων.

Το Berkeley NOW (Network of Workstations), το σύστημα συλλέγει δεδομένα διασποράς και τα αποθηκεύει σε μια βάση δεδομένων, ενώ ένα σύστημα όπως το Parmon, που αναπτύχθηκε στην Ινδία, επιτρέπει την οπτική παρατήρηση και τη διαχείριση των μεγάλων cluster. [55]

Η εφαρμογή Checkpointing μπορεί να χρησιμοποιηθεί για να αποκαταστήσει μια δεδομένη κατάσταση του συστήματος, όταν ένας κόμβος αποτύχει κατά τη διάρκεια μιας μακράς περιόδου υπολογισμού σε πολλαπλούς κόμβους. Αυτό είναι σημαντικό σε μεγάλα cluster, δεδομένου ότι, καθώς ο αριθμός των κόμβων αυξάνεται, το ίδιο συμβαίνει και με τη πιθανότητα του «προβληματικού» κόμβου κάτω από βαριά φορτία υπολογισμού.[57] Σημεία Ελέγχου (Checkpoint) μπορούν να επαναφέρουν το σύστημα σε μια σταθερή κατάσταση έτσι ώστε η επεξεργασία μπορεί να ξαναρχίσει χωρίς να χρειάζεται να υπολογίσει εκ νέου τα αποτελέσματα.[57]

2.10 Άλλες προσεγγίσεις

Αν και τα περισσότερα computer cluster είναι μόνιμα εξαρτήματα, απόπειρες flash mob computing έχουν γίνει για τη δημιουργία βραχύβιων clusters για συγκεκριμένους υπολογισμούς. Ωστόσο, μεγαλύτερη κλίμακα «εθελοντών» υπολογιστικών συστημάτων, όπως τα συστήματα που βασίζονται σε BOINC έχουν περισσότερους οπαδούς.

2.10.1 Flash mob computing

Flash mob computing (ή flash mob computer) είναι ένα προσωρινό ad hoc computer cluster που τρέχει ειδικό λογισμικό για να συντονίσει τους επιμέρους υπολογιστές σε ένα ενιαίο υπερυπολογιστή. Ένας υπολογιστής flash mob είναι διαφορετικός από τα άλλα είδη των cluster computers σε ότι έχει συσταθεί και «διαλυθεί» την ίδια ημέρα ή κατά τη διάρκεια, σε σύντομο χρονικό διάστημα και περιλαμβάνει πολλούς ανεξάρτητους ιδιοκτήτες ηλεκτρονικών

υπολογιστών που έρχονται μαζί σε μια κεντρική φυσική θέση να εργαστούν σε ένα συγκεκριμένο πρόβλημα ή και κοινωνική εκδήλωση.

Ο Flash mob υπολογιστής αντλεί το όνομά του από την πιο γενική έννοια flash mob που μπορεί να σημαίνει κάθε δραστηριότητα που περιλαμβάνει πολλούς ανθρώπους που συντονίστηκαν μέσω των εικονικών κοινοτήτων που έρχονται μαζί για σύντομο χρονικό διάστημα για μια συγκεκριμένη εργασία ή συμβάν. Το Flash mob computing είναι ένας πιο συγκεκριμένος τύπος του flash mob με σκοπό την προσέγγιση των ανθρώπων και των υπολογιστών τους μαζί για να εργαστούν σε ένα ενιαίο έργο ή συμβάν.

Ο πρώτος υπολογιστής flash mob δημιουργήθηκε στις 3 Απριλίου 2004 στο Πανεπιστήμιο του Σαν Φρανσίσκο, χρησιμοποιώντας λογισμικό γραμμένο σε USF και ονομάζεται FlashMob (δεν πρέπει να συγχέεται με το γενικότερο όρο flash mob). Η εκδήλωση, που ονομάζεται FlashMob I, ήταν μια επιτυχία. Υπήρξε μια πρόσκληση για υπολογιστές στην ειδησεογραφική ιστοσελίδα για υπολογιστές Slashdot. Ένα άρθρο της NY Times "Hey, Gang, Let's Make Our Own Supercomputer" («ελάτε να φτιάξουμε τον δικό μας υπερυπολογιστή») έφερε πολύ προσοχή στην προσπάθεια. Περισσότεροι από 700 ηλεκτρονικοί υπολογιστές ήρθαν στο γυμναστήριο του Πανεπιστημίου του Σαν Φρανσίσκο και ήταν συνδεδεμένοι σε δίκτυο που δώρισε η Foundry Networks. Στο FlashNob I ήταν σε θέση να τρέξει ως ένα σημείο αναφοράς για 256 υπολογιστές και πέτυχε μια μέγιστη απόδοση των 180 Gflops (δισεκατομμύρια υπολογισμούς ανά δευτερόλεπτο), ενώ ο υπολογισμός αυτός σταμάτησε τα τρία τέταρτα της διαδρομής και αυτό οφείλεται σε μια αποτυχία ενός κόμβου. Η καλίρερη πλήρες λειτουργία χρησιμοποιείσαι 150 υπολογιστές και είχε ως αποτέλεσμα το 77 Gflops. Το Flashmob έτρεχε από ένα bootable CD-ROM που έτρεχε ένα αντίγραφο του Morphix Linux και ήταν διαθέσιμο μόνο για την πλατφόρμα x86.

Παρά τις προσπάθειες αυτές, το έργο δεν ήταν σε θέση να επιτύχει τον αρχικό στόχο της τρέχει ένα cluster στιγμιαία, αρκετά γρήγορα για να εισέλθουν στον (Νοέμβριος 2003) κατάλογο των 500 κορυφαίων υπερυπολογιστών. Το σύστημα θα έπρεπε να παρέχει τουλάχιστον 402,5 Gflops για να αντιστοιχεί με ένα κινέζικο Cluster των 256 κόμβων με Intel Xeon CPU . [58]

Εικόνα: 2.2 Flash Mob computing από 256 εθελοντές με τους προσωπικούς τους υπολογιστές, Laptop & Desktop συνέθεσαν ένα cluster computer στο πανεπιστήμιο του San Francisco.

2.11 Λειτουργικά Συστήματα για Computer Cluster

Υπάρχουν αρκετά είδη Λειτουργικών Συστημάτων που είναι κατάλληλα διαμορφωμένα και φέρουν τα κατάλληλα εργαλεία και σουίτες για δημιουργία και διαχείριση του εκάστοτε Cluster. Τα περισσότερα από αυτά βασίζονται σε εκδόσεις των Linux μιας και είναι ένα Λειτουργικό σύστημα ανοιχτού κώδικα (open source) και είναι ελεύθερη η διανομή του, δηλαδή χωρίς κάποια χρέωση για αγορά ή για συντήρηση, update κλπ. και μπορεί ο κάθε προγραμματιστής να το διαμορφώσει ανάλογα με τις ανάγκες του. Επίσης υπάρχουν και Λειτουργικά Συστήματα «έτυμα» διαμορφωμένα από εταιρίες όπως είναι η Microsoft με διάφορες διανομές Windows Server κατάλληλα διαμορφωμένες, επίσης η Oracle με το Solaris όπου πλέον συγχωνεύθηκε με την Sun Microsystems και η Apple με το Xgrid για για δημιουργία και διαχείριση cluster.

2.11.1 Red Hat Cluster suite

Το Red Hat Cluster περιλαμβάνει λογισμικό για να δημιουργεί Cluster υψηλής διαθεσιμότητας (High Availiability Cluster) και Cluster κατανομής φορτίου (load balancing cluster). Και τα δύο μπορούν να χρησιμοποιηθούν στο ίδιο σύστημα, αν και αυτή η περίπτωση χρήσης είναι απίθανη. Και τα δύο προϊόντα, το High Availiability Add-0n και το Load Balancer

Add-On, βασίζονται στη κοινότητα έργων Open Source. Οι προγραμματιστές του Red Hat Cluster «ανεβάζουν» και δίνουν τον κώδικα στην κοινότητα. [59][60]

2.11.1.1 High Availability Add-on

Το High Availability Add-On είναι η εφαρμογή του Red Hat Linux από-HA. Επιχειρεί να εξασφαλιστεί η διαθεσιμότητα των υπηρεσιών παρακολουθώντας άλλους κόμβους του cluster. Όλοι οι κόμβοι του cluster πρέπει να συμφωνήσουν σχετικά με τη διαμόρφωση και την κατάσταση των κοινών υπηρεσιών τους, πριν θεωρηθεί ότι το Cluster βρίσκεται σε απαρτία και οι υπηρεσίες είναι σε θέση να ξεκινήσουν. Η κύρια μορφή επικοινωνίας για τη κατάσταση του κόμβου είναι μέσω μιας συσκευής δικτύου (συνήθως Ethernet), αν και στην περίπτωση πιθανής βλάβης του δικτύου, μπορεί να αποφασιστεί μέσω άλλης μεθόδου, όπως η κατανεμημένη αποθήκευση (shared storage) ή multicast. Υπηρεσίες λογισμικού, συστήματα αρχείων και η κατάσταση του δικτύου μπορεί να παρακολουθείται και να ελέγχεται από την σουίτα cluster, υπηρεσιών και πόρων μπορεί να αποτύχει πάνω σε άλλους κόμβους του δικτύου, σε περίπτωση αποτυχίας. Το Cluster τερματίζει βίαια τη πρόσβαση ενός κόμβου του cluster σε υπηρεσίες ή πόρους, μέσω «περίφραξης», για να εξασφαλίσει τον κόμβο και τα δεδομένα σε μια γνωστή κατάσταση. Ο κόμβος τερματίζεται με την αφαίρεση «εξουσίας» (γνωστή ως STONITH) ή πρόσβαση στην κοινόχρηστη αποθήκη-δεξαμενή δεδομένων. Η Υπηρεσία κλειδώματος και ελέγχου είναι εγγυημένη μέσω της «περίφραξης» και STONITH.

Πιο πρόσφατες εκδόσεις του Red Hat χρησιμοποιούν ένα κατανεμημένο σύστημα διαχειριστής ασφάλειας, που επιτρέπει επιλεγμένο κλείδωμα ασφάλισης και όχι ένα ενιαίο σημείο της αποτυχίας. Οι παλαιότερες εκδόσεις της σουίτας cluster στηρίχθηκε σε ένα «μεγάλο ενοποιημένο διαχειριστή κλειδώματος» (GULM) που θα μπορούσαν να ομαδοποιηθούν, αλλά εξακολουθεί να παρουσιάζει ανακατεύθυνση (failover) αν οι κόμβοι λειτουργούν ως διακομιστές GULM και αποτύχουν. Το GULM ήταν τελευταία φορά διαθέσιμο στο Red Hat Cluster Suite 4.

2.11.1.2 Τεχνικές λεπτομέρειες

- Υποστηρίζει μέχρι 128 κόμβους

- NFS (Unix) /CIFS/GFS/GFS2 (Multiple Operating Systems) Υποστήριξη σε Περίπτωση αποτυχίας του συστήματος αρχείων.

- Υποστήριξη υπηρεσίας σε σημείου επαναφοράς σε περίπτωση αποτυχίας.

- Πλήρες διαμοιρασμένο υποσύστημα αποθήκευσης δεδομένων.

- Εγγυάται την ακεραιότητα των δεδομένων.

- SCSI και κανάλι οπτικών ινών.

- OCF και LSM παράγοντες πόρων.

2.11.1.3 Add-on Κατανομής Φορτίου

Το Red Hat υιοθέτησε το λογισμικό κατανομής φορτίου Piranha για να καταστεί δυνατή η διαφανής κατανομή φορτίου και failover διακομιστές. Η εφαρμογή έχει δικιά της βάση δεν απαιτεί ειδική διαμόρφωση για να είναι ισορροπημένη, αντί ένός διακομιστής Red Hat Enterprise Linux ρυθμισμένος για κατανομή φορτίου, παρακολουθεί τους διάδρομους κυκλοφορίας και βασίζεται σε μετρήσεις / κανόνες για τα εφαρμόζει στην κατανομή φορτίου.

2.11.1.4 Υποστήριξη και Κύκλος ζωής

Το Red Hat Cluster suite είναι συνδεδεμένο με μια αντίστοιχη έκδοση του Red Hat Enterprise Linux και ακολουθεί την ίδια πολιτική συντήρησης. Το προϊόν δεν έχει καμία ενεργοποίηση, προθεσμία ή remote kill switch και μπορεί να παραμείνει να λειτουργεί μετά την λήξη του κύκλου ζωής υποστήριξης. Είναι εν μέρει υποστηρίζεται να «τρέχει» κάτω από VMware Virtual Machine (εικονικά).[61]

2.11.2 Microsoft Cluster Server

Το Microsoft Cluster Server (MSCS) είναι ένα πρόγραμμα υπολογιστή που επιτρέπει στους υπολογιστές Server να εργάζονται μαζί ως ένα computer cluster, να παρέχει failover (ανακατεύθυνση) και να αυξάνει την διαθεσιμότητα των εφαρμογών ή τη παράλληλη δυνατότητα υπολογισμού στην περίπτωση των υπολογιστών Clusters υψηλής απόδοσης (HPC) (όπως σε υπερυπολογιστές).

Η Microsoft έχει τρεις τεχνολογίες για Clustering: Microsoft Cluster Service (MSCS), Component Load Balancing (OEB) (μέρος του Application Center 2000), και το Nerwork Load

Balancing (NLB). Στα Windows Server 2008 και Windows Server 2008 R2 η υπηρεσία MSCS έχει μετονομαστεί σε Windows Server Failover Clustering και το Component Load Balancing (OEB) χαρακτηριστικό έχει αφαιρεθεί.

2.11.2.1 Υποστήριξη

Το Cluster Server είχε την κωδική ονομασία «Wolfpack" κατά την ανάπτυξή του.[62] Τα Windows NT Server 4.0 Enterprise Edition ήταν η πρώτη έκδοση των Windows η οποία συμπεριλάβανε το λογισμικό MSCS (Microsoft Cluster Service). Το λογισμικό έχει έκτοτε ανανεώνεται (updated) με κάθε νέα έκδοση του διακομιστή. Το λογισμικό του cluster αξιολογεί τους πόρους των servers στο cluster και επιλέγει ποιοι θα χρησιμοποιηθούν με βάση κριτήρια που καθορίζονται στη μονάδα διαχείρισης. Τον Ιούνιο του 2006, η Microsoft κυκλοφόρησε τα Windows Compute Cluster Server 2003,[63] η πρώτη τεχνολογία cluster υπολογιστών υψηλής απόδοσης (HPC) από τη Microsoft. Η ποιο πρόσφατη έκδοση της Microsoft για clustering είναι αυτή των Windows Server 2012R2.s

2.11.3 Solaris Cluster

Το Solaris Cluster (μερικές φορές Sun Cluster ή SunCluster) είναι ένα προϊόν λογισμικού για υψηλής διαθεσιμότητας cluster (High Availability Clusters) για το λειτουργικό σύστημα Solaris, που αρχικά δημιουργήθηκε από τη Sun Microsystems, η οποία εξαγοράστηκε από την Oracle Corporation το 2010 και χρησιμοποιείται για να βελτιωθεί η διαθεσιμότητα των υπηρεσιών λογισμικού, όπως βάσεις δεδομένων, κοινή χρήση αρχείων σε ένα δίκτυο, το ηλεκτρονικό εμπόριο ιστοσελίδες, ή άλλες εφαρμογές. Το Sun Cluster λειτουργεί έχοντας περιττούς υπολογιστές ή κόμβους, όπου ένα ή περισσότεροι υπολογιστές να συνεχίσουν να παρέχουν την εκάστοτε υπηρεσία, αν ένα άλλος αποτύχει. Κόμβοι μπορούν να βρίσκονται στο ίδιο κέντρο δεδομένων ή και σε διαφορετικές ηπείρους.

Εικόνα: 2.3 Sun Microsystems Solaris Cluster

2.11.3.1 Υποστήριξη και Χαρακτηριστικά

Το Solaris Cluster παρέχει τις υπηρεσίες που παραμένουν διαθέσιμες ακόμη και όταν μεμονωμένοι κόμβοι ή στοιχεία του cluster αποτύχουν. Το Solaris Cluster παρέχει δύο τύπους υπηρεσιών HA: Υπηρεσίες ανακατεύθυνσης (failover services) και κλιμακούμενες υπηρεσίες (scalable services). Για την εξάλειψη των ενιαίων σημείων σε περίπτωση αποτυχίας, η διαμόρφωση Solaris Cluster διαθέτει εφεδρικά εξαρτήματα και χαρακτηρίστηκα, συμπεριλαμβανομένων των πολλαπλών συνδέσεων του δικτύου και αποθήκευσης των δεδομένων, τα οποία πολλαπλασιάζονται συνδεόμενα μέσω ενός δικτύου. σημείου-περιοχής αποθήκευσης (SAN-Storage Area Network). Το λογισμικό clustering, όπως το Solaris Cluster είναι ένα βασικό συστατικό, κλειδί σε μια λύση Business Continuity και το Solaris Cluster Geographic Edition δημιουργήθηκε ειδικά για να αντιμετωπίσει αυτή την απαίτηση.

Το Solaris Cluster είναι ένα παράδειγμα του επιπέδου kernel λογισμικού clustering. Ορισμένες από τις διεργασίες που «τρέχει» είναι κανονικές διεργασίες του συστήματος για τα συστήματα που λειτουργεί, αλλά έχει και κάποια ειδική πρόσβαση στο λειτουργικό σύστημα και τις λειτουργίες του πυρήνα για συστήματα που φιλοξενεί. Τον Ιούνιο του 2007, η Sun κυκλοφόρησε τον πηγαίο κώδικα του Solaris Cluster μέσω της κοινότητας OpenSolaris HA Clusters.[64]

2.11.3.2 Έκδοση Solaris Cluster Geographic

Το SCGE είναι ένα πλαίσιο διαχείρισης που εισήχθη τον Αύγουστο του 2005. Επιτρέπει δύο εγκαταστάσεις Solaris Cluster όπου αντιμετωπίζεται ως ενιαία, σε συνδυασμό με ένα ή περισσότερα προϊόντα αντιγραφής δεδομένων, να παρέχει Disaster Recovery για μία εγκατάσταση σε ηλεκτρονικό υπολογιστή. Με την εξασφάλιση ότι οι ενημερώσεις δεδομένων (updates) είναι συνεχές και αναπαράγονται σε μια απομακρυσμένη τοποθεσία σε σχεδόν πραγματικό χρόνο (near-real time), η τοποθεσία μπορεί να πάρει γρήγορα ολόκληρη την παροχή μιας υπηρεσίας σε περίπτωση που το σύνολο της πρώτης περιοχής χαθεί ως αποτέλεσμα μιας καταστροφής, είτε από φυσικό είτε ανθρωπογενή παράγοντα. Αυτό είναι το κλειδί για την ελαχιστοποίηση του σημείου δεδομένων ανάκτησης (RPO-Revovery Point Object) και του χρόνου ανάκτησης δεδομένων (RTO-Recovery Time Objective) για την υπηρεσία.

2.11.3.3 Proxy File System

Το PxFS (σύστημα αρχείων Proxy) είναι κατανεμημένο, υψηλής διαθεσιμότητα (High Availability). Το POSIX είναι συμβατό με το εσωτερικό σύστημα αρχείων των κόμβων του Solaris Cluster. Οι συσκευές που χρησιμοποιούνται Sun Cluster παγκοσμίως είναι πιθανόν στηριγμένες στο PxFS (Proxy File System).[65]

2.11.3.4 Υποστηριζόμενες εφαρμογές

Το Solaris Cluster χρησιμοποιεί στοιχεία λογισμικού που ονομάζεται agents «πράκτορες», οι οποίοι παρακολουθούν μια εφαρμογή για να ανιχνεύσουν αν λειτουργεί σωστά, και να αναλάβουν δράση, εάν εντοπιστεί κάποιο πρόβλημα. Επίσης «πράκτορες» για κοινές εφαρμογές που περιλαμβάνονται, όπως Siebel Systems, SAP LiveCache, WebLogic Server, Sun Java Application Server, MySQL, Oracle RAC, Oracle E-Business Suite και Samba, μεταξύ άλλων υπάρχει επίσης και ένας οδηγός που επιτρέπει την εκτελεστής στο cluster για να την δημιουργία «πρακτόρων» για άλλες εφαρμογές**.**

2.11.4 Apple Xgrid

Το Xgrid είναι ένα ιδιόκτητο πρόγραμμα και χρησιμοποιεί κατανεμημένο πρωτόκολλο υπολογιστών που αναπτύχθηκε από την Advanced Computation Group, θυγατρική της Apple Inc που επιτρέπει δικτυωμένους υπολογιστές να συμβάλουν σε ένα ενιαίο έργο.

Παρέχει στους διαχειριστές δικτύων μια μέθοδο για τη δημιουργία ενός cluster computing, το οποίο τους επιτρέπει να εκμεταλλεύονται την περισσευούμενη υπολογιστική ισχύ για άλλους υπολογισμούς καθώς μπορεί να χωρίσει την ισχύ εύκολα και να τη διαθέσει σε μικρότερες διεργασίες, όπως Mandelbrot χάρτες. Η εγκατάσταση ενός cluster Xgrid μπορεί να επιτευχθεί με πολύ μικρό κόστος. Σαν Xgrid client (πελάτης) είναι προεγκατεστημένο σε όλους τους υπολογιστές που τρέχουν Mac OS X 10.4 με το Mac OS X 10.7. Ο client (πελάτης) Xgrid δεν περιλαμβάνεται στο Mac OS X 10.8. Ο ελεγκτής Xgrid, ο προγραμματιστής διεργασιών της λειτουργίας Xgrid, επίσης περιλαμβάνεται στο Mac OS X Server και ως δωρεάν download από την Apple. Η Apple έχει κρατήσει τη γραμμή εντολών (comand line) για ελέγχω των διεργασιών, έναν μηχανισμό αρκετά μινιμαλιστικό, ενώ παρέχει ένα API για την ανάπτυξη πιο εξελιγμένων εργαλείων χτισμένα γύρω από αυτό.

Το πρόγραμμα χρησιμοποιεί το δικό του πρωτόκολλο επικοινωνίας σε ένα κορυφαίο επίπεδο ώστε να επικοινωνεί με τους άλλους κόμβους. Αυτό το πρωτόκολλο επικοινωνίας

διεπαφών με την υποδομή BEEP, ένα πλαίσιο πρωτόκολλου δικτύου εφαρμογής. Οι Υπολογιστές εντοπίζονται από το σύστημα Xgrid, δηλαδή υπολογιστές με την υπηρεσία του Mac OS X Xgrid ενεργοποιημένη, προστίθενται αυτόματα στη λίστα με τους διαθέσιμους υπολογιστές για να χρησιμοποιούν τις διεργασίες προς επεξεργασία.

Όταν ο αρχικός υπολογιστής στέλνει τις πλήρεις οδηγίες ή εργασία για την επεξεργασία στον ελεγκτή, ο ελεγκτής χωρίζει την εργασία σε μικρά πακέτα οδηγιών, που είναι γνωστά ως διεργασίες. Ο σχεδιασμός του συστήματος Xgrid αποτελείται από αυτά τα μικρά πακέτα τα οποία μεταφέρονται σε όλους τους Xgrid-ενεργοποιημένους υπολογιστές στο δίκτυο. Αυτοί οι υπολογιστές ή κόμβοι, εκτελούν τις οδηγίες που παρέχονται από τον ελεγκτή και στη συνέχεια επιστρέφουν τα αποτελέσματα. Ο ελεγκτής συγκεντρώνει τα επιμέρους αποτελέσματα εργασιών σε ολόκληρα αποτελέσματα εργασίας και τα επιστρέφει στον αρχικό υπολογιστή.

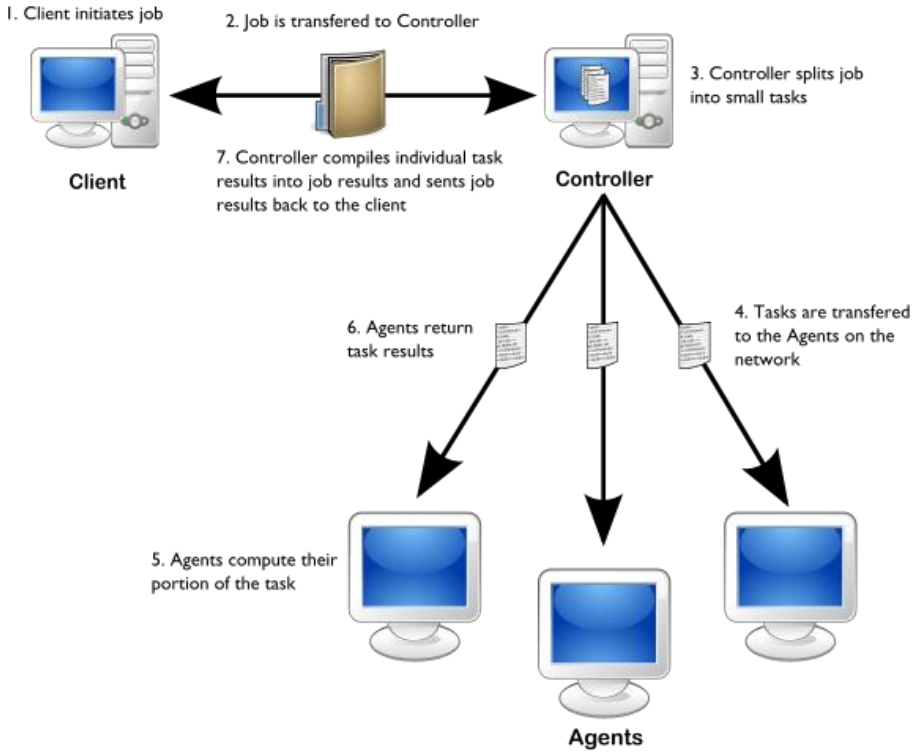
Η Apple διαμόρφωσε το σχεδιασμό της Xgrid με τη βοήθεια του προγράμματος Zilla, κατανεμημένο με το Λειτουργικό Σύστημα NeXT's OPENSTEP διεπαφή προγραμματισμού εφαρμογών (API), το οποίο η Apple κατέχει τα δικαιώματα. Η εταιρεία επέλεξε επίσης να παρέχουν την έκδοση του client του Mac OS X με λειτουργίες μόνο γραμμής εντολών και μικρή ευελιξία, δίνοντας παράλληλα στην έκδοση του Mac OS X Server του Xgrid ένα πίνακα ελέγχου GUI και ένα πλήρες σύνολο χαρακτηριστικών.

2.11.4.1 Πρωτόκολλο

Το πρωτόκολλο χρησιμοποιεί το πλαίσιο του δικτύου BEEP για να επικοινωνεί με τους κόμβους στο δίκτυο. Η υποδομή του συστήματος περιλαμβάνει τρεις τύπους υπολογιστών που επικοινωνούν μέσω του πρωτοκόλλου. Ένας είναι ο πελάτης (client), ο οποίος επικοινωνεί με τον υπολογισμό. Επόμενος είναι ο ελεγκτής (controler), ο οποίος ξεκινά και απομονώνει τον υπολογισμό. Τέλος, οι «πράκτορες» που επεξεργάζονται το δικό κομμάτι που τους έχει διατεθεί μέρος του υπολογισμού.

Ένας υπολογιστής μπορεί να κάνει χρήση ενός ή και των τριών από αυτά τα χαρακτηριστικά την ίδια στιγμή. Το πρωτόκολλο Xgrid παρέχει τη βασική υποδομή για τους υπολογιστές για να επικοινωνούν, αλλά δεν συμμετέχει στην επεξεργασία του συγκεκριμένου υπολογισμού.[66] Το Xgrid απευθύνεται σε χρονοβόρους υπολογισμούς που μπορεί εύκολα να διαχωρίζονται-διασπώνται σε μικρότερες εργασίες, μερικές φορές ονομάζονται παράλληλες διεργασίες.[67] Αυτό περιλαμβάνει τους υπολογισμούς Monte Carlo, 3D rendering και Mandelbrot χάρτες.[66]

Στο πλαίσιο του πρωτοκόλλου Xgrid, τρεις τύποι μηνυμάτων μπορεί να περάσουν σε άλλους υπολογιστές στο ίδιο cluster: αιτήματα (request), κοινοποιήσεις και οι απαντήσεις. Στις αιτήσεις θα πρέπει να ανταποκρίνονται στον παραλήπτη με μία απάντηση, οι κοινοποιήσεις δεν απαιτούν μια απάντηση και οι απαντήσεις



Σχήμα: 2.9 Xgrid Protocol

ανταποκρίνονται για τα απεσταλμένα μηνύματα. Αυτά προσδιορίζονται από το όνομά τους, τον τύπο (αίτημα / γνωστοποίηση / απάντηση) και το περιεχόμενο. Κάθε μήνυμα είναι έγκλειστο σε ένα μήνυμα BEEP (BEEP MSG) και αναγνωρίζεται από την παραλαβή από μια άδεια απάντηση (RPY).[68] Το Xgrid δεν κατασκευάζει την δομή των BEEP μηνυμάτων/ απαντήσεων. Κάθε μήνυμα που λαμβάνετε και απαιτεί μια απάντηση απλώς δημιουργεί ένα ανεξάρτητο μήνυμα BEEP που περιέχει την απάντηση. Τα μηνύματα Xgrid κωδικοποιούνται ως λεξικά ζεύγη κλειδιού / τιμής (key/value) που μετατρέπονται σε XML προτού αποσταλεί στο δίκτυο BEEP.

2.11.4.2 Αρχιτεκτονική

Η αρχιτεκτονική του συστήματος Xgrid έχει σχεδιαστεί γύρω από ένα σύστημα με βάση τις διεργασίες. Ο ελεγκτής στέλνει «πράκτορες» (agents) με εργασίες, και οι «πράκτορες» επιστρέφουν τις απαντήσεις (responses). Τον πραγματικό υπολογισμό, όπου ο ελεγκτής εκτελεί σε ένα σύστημα Xgrid είναι γνωστό ως μια εργασία. Η εργασία περιέχει όλα τα αρχεία που

απαιτούνται για να ολοκληρωθεί το έργο με επιτυχία, όπως τις παραμέτρους εισόδου, αρχεία δεδομένων, καταλόγους, εκτελέσιμα αρχεία ή και προγράμματα κελύφους (Shell scripts), τα αρχεία που περιλαμβάνονται σε μια δουλειά Xgrid πρέπει να είναι σε θέση να εκτελούνται είτε ταυτόχρονα είτε ασύγχρονα, ή τυχόν οφέλη από τη λειτουργία μιας τέτοιας εργασίας σε Xgrid μπορεί να χαθεί. Όταν η εργασία ολοκληρωθεί, ο ελεγκτής μπορεί να ρυθμιστεί ώστε να ενημερώσει τον πελάτη (client) για την ολοκλήρωση ή την αποτυχία της αποστολής του, για παράδειγμα μέσω email. Ο πελάτης μπορεί να αφήσει το δίκτυο, ενώ οι εργασίες εκτελούνται. Μπορεί επίσης να παρακολουθεί (monitoring) την κατάσταση της εργασίας με αίτημα η από τον πίνακα του ελεγκτή, αν και δεν μπορεί να παρακολουθεί τη συνεχιζόμενη πρόοδο των επιμέρους εργασιών.[69] Ο ελεγκτής είναι κεντρικής σημασίας για τη σωστή λειτουργία ενός Xgrid, όπως ο κόμβος (node) αυτός είναι υπεύθυνος για τη διανομή, την εποπτεία και τον συντονισμό των εργασιών για τους «πράκτορες» (agents).

Το πρόγραμμα που τρέχει στον ελεγκτή μπορεί να εκχωρήσει και τον επαναπροσδιορισμό των καθηκόντων για να χειριστεί μεμονωμένες αποτυχίες agents κατά αίτημα. Ο αριθμός των καθηκόντων που ανατίθενται σε έναν agent εξαρτάται από δύο παράγοντες: τον αριθμό των agents σε ένα Xgrid και τον αριθμό των επεξεργαστών σε κάθε κόμβο. Ο αριθμός των «πρακτόρων» (agents) σε ένα Xgrid καθορίζει το πώς ο ελεγκτής θα αναθέσει καθήκοντα. Τα καθήκοντα που μπορούν να ανατεθούν ταυτοχρόνως για μεγάλο αριθμό «πρακτόρων» ή στην ουρά για ένα μικρό αριθμό «πρακτόρων». Όταν ένας κόμβος (node) με περισσότερους από έναν επεξεργαστή ανιχνεύεται σε Xgrid, ο ελεγκτής (controler) μπορεί να αναθέσει μία εργασία ανά επεξεργαστή. Αυτό συμβαίνει μόνο εάν ο αριθμός των agents στο δίκτυο είναι μικρότερος από τον αριθμό των εργασιών τότε ο ελεγκτής πρέπει να ολοκληρώσει.[69]

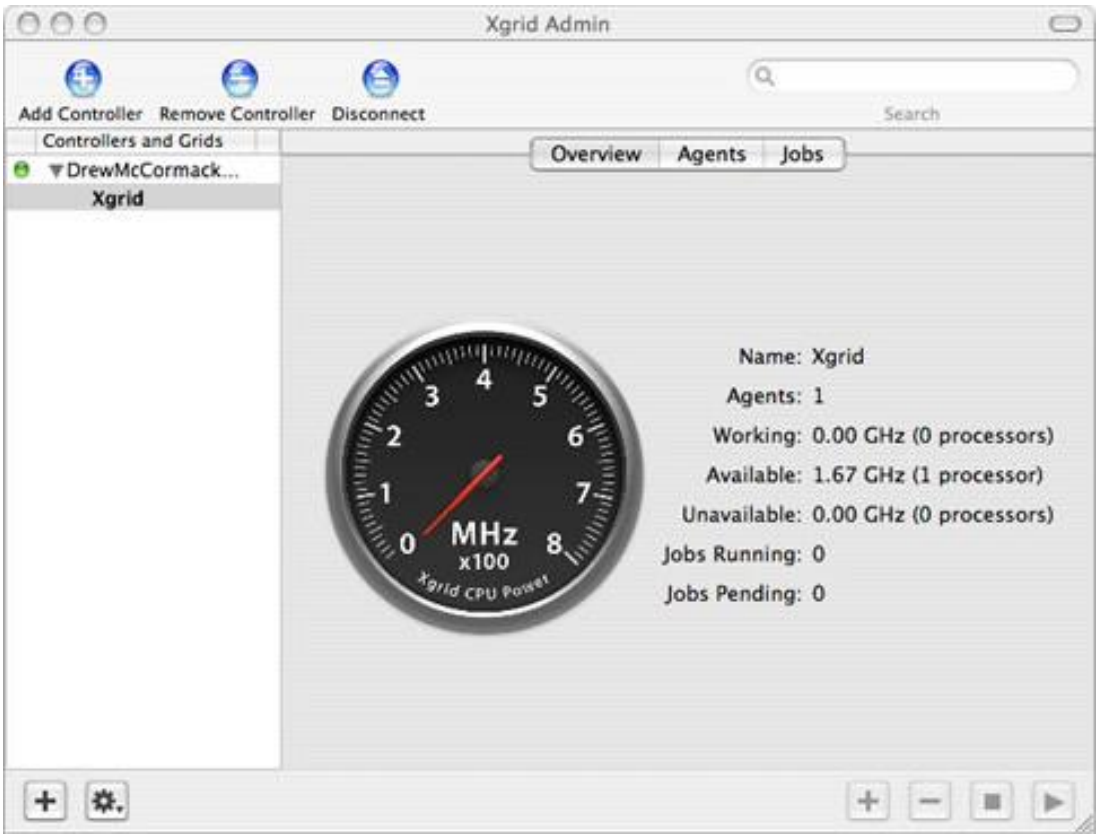
Το Xgrid στρωματοποιείται επί του Block Extensible Exchange Protocol (BEEP), ένα πρότυπο IETF συγκρίσιμο με HTTP, αλλά με έμφαση στην αμφίδρομη επικοινωνία πολυπλεξίας, όπως αυτό που συμβένει σε peer-to-peer δίκτυα. Το BEEP, με τη σειρά του χρησιμοποιεί XML για να καθορίσει τα προφίλ για την επικοινωνία μεταξύ πολλαπλών παραγόντων μέσω ενός ενιαίου δικτύου ή σύνδεση στο internet.[70]

2.11.4.3 Διεπαφή-Interface

Ενώ είναι δυνατή η πρόσβαση Xgrid από τη γραμμή εντολών (comand line), η γραφική διεπαφή χρήστη Xgrid, σε ένα πρόγραμμα συνδυασμένο με το Mac OS X Server ως τον Μάρτιο του 2009, οπου διατίθεται online , είναι ένας πολύ πιο αποτελεσματικός τρόπος για τη

διαχείριση ενός συστήματος Xgrid. Αρχικά, ο agent Xgrid είχε συμπεριληφθεί και εγκατασταθεί σε όλα τα Mac OS X έκδοση 10.4 , αλλά το GUI (Graphical User Interface) ήταν αποκλειστικά για τους χρήστες του Mac OS X Server. Αυτή η απόφαση περιόρισε τις προσπάθειες της κοινότητας και κατόχων υπολογιστή Mac για να αγκαλιάσουν την πλατφόρμα. Τελικά, η Apple κυκλοφόρησε το Mac OS X Server Administrator Tools για το κοινό, το οποίο περιελάβανε την εφαρμογή διαχείρισης Xgrid πακέτο με το Mac OS X Server.[71]

Παρά την έλλειψη ενός γραφικού περιβάλλοντος ελέγχου στο πρότυπο (non-server), Mac OS X διανομής, είναι δυνατόν να δημιουργήσει ένα ελεγκτή Xgrid μέσω των εργαλείων της γραμμής εντολών xgridctl και xgrid. Όταν ο ελεγκτής Xgrid daemon τρέχει, h διαχείριση του δικτύου με το εργαλείο Xgrid Admin της Apple είναι δυνατή.[71] Ορισμένες εφαρμογές, όπως VisualHub, παρέχουν τη δυνατότητα ελέγχου Xgrid μέσω διεπαφών των χρηστών τους.[72][73]



Εικόνα: 2.4 Xgrid Administration Tool (Εργαλείο Διαχείρισης)

Αναφορές και Βιβλιογραφία

Πρωτεύουσες Αναφορές :

- [1]Computer Cluster (n.d)In Wikipedia. Retrieved 18-09-2013 https://en.wikipedia.org /wiki/ Computer_cluster,
- [2]Supercomputer(n.d) In Wikipedia. Retrieved 20-09-2013 https://en.wikipedia.org/wiki /Supercomputer.
- [3]Contractors-computer (n.d) In Railway-Technology . Retrieved 14-08-2014. http://www.railway-technology.com/contractors/computer/men-mikro/
- [4] Load Balancing Computing (n.d)In Wikipedia. Retrieved 20-11-2013. https://en.wikipedia.org/wiki/Load_balancing_(computing).
- [5]DNS-Round Robin(n.d)In thetechnologychronicle.blogspot.in dns-round-robin 05-11-2013
- [6] IEEE-802.1aq (08-05-2012) In Wikipedia. Retrieved 02-06-2013. https://en.wikipedia .org/wiki/IEEE_802.1aq.
- [7] IEEE-802.1aq (24-02-2011) In Wikipedia. Retrieved 11-05-2013. https://en.wikipedia .org/wiki/IEEE_802.1aq.
- [8] Network Topology (11-05-2012) In Wikipedia. Retrieved 11-05-2013 https://en.wikipedia.org/wiki/Network_topology.
- [9] IEEE-802.1aq Shortest Path (07-05-2012)In Wikipedia. Retrieved 11-05-2013. https://en.wikipedia.org/wiki/ IEEE_802.1aq.
- [10] High Performance Computing Science (11-12-2005) In Johroline. Retrieved 12-05-2013 http://www.johronline.com/issue/20131212-212933.063.pdf
- [11] Degima (n.d)In Wikipedia. Retrieved 15-05-2013. https://en.wikipedia.org https://en.wikipedia.org/wiki/DEGIMA
- [12] Multiple-Parallel-Algorithm (03-05-2009) In Cct.Isu.edu. Retrieved 07-07-2014 https://www.cct.lsu.edu/~korobkin/tmp/SC10/papers/pdfs/gb106s4.pdf
- [13]Multiple-Parallel-Algorithm (03-05-2009) In Cs.bris.ac.uk Retrieved 07-07-2014 http://www.cs.bris.ac.uk/~simonm/conferences/isc09/hamada.pdf
- [14] PlayStation3 Cluster (08-05-2012) In Wikipedia. Retrieved 03-08-2014 https://en.wikipedia.org/wiki/PlayStation_3_cluster.
- [15] PlayStation3 Cluster (08-05-2012) In Wikipedia. Retrieved 03-08-2014 https://en.wikipedia.org/wiki/PlayStation_3_cluster.
- [16] Yellow dog Linux on Sony PlayStation3 (08-04-2012) In Akdavetaylor. Retrieved 12-05-2014 http://www.askdavetaylor.com/yellow_dog_linux_on_sony_playstation3/
- [17] Sony PlayStation3 Cluster (05-06-2012) In Wikipedia. Retrieved 12-05-2014. https://en.wikipedia.org/wiki/PlayStation_3_cluster.
- [18]Yellow Dog Linux (10-06-2012) In Wikipedia. Retrieved 07-05-2014. https://en.wikipedia.org/wiki/Yellow_Dog_Linux.
- [19] Academic PlayStation3 Computing Cluster (09-08-2012)In Phys.org-news. Retrieved 08-06-2014. http://phys.org/news/2007-03-academic-playstation-cluster.html.

- [20] Academic PlayStation Cluster (10-07-2012) In Phys.org-news. Retrieved 08-06-2014 http://phys.org/news/2007-03-academic-playstation-cluster.html.
- [21] Academic PlayStation Cluster. (10-07-2012) In Phys.org-news. Retrieved 08-06-2014 http://phys.org/news/2007-03-academic-playstation-cluster.html.
- [22] PS3 Gravity Grid (30-10-2009) In Spectre Group. Wordpress. Retrieved 10-06-2014 https://spectregroup.wordpress.com/2009/10/30/make-your-own-supercompute/.
- [23] Supercomputer (17-10-2010) In Wired-News. Retrieved 12-07-2013. http://archive .wired.com/techbiz/it/news/2007/10/ps3_supercomputer
- [24] Homemade Cluster (16-08-2011)In Computerworld. Retrieved 12-07-2013. http://www.computerworld.com/article/2539437/high-performance-computing/ps3-cluster-creates-homemade--cheaper-supercomputer.html
- [25] PlayStation3 Cluster (17-02-17) In Wikipedia. Retrieved 23-06-2014. https://en.wikipedia.org/wiki/PlayStation_3_cluster.
- [26] PlayStation3 Cluster (23-12-2008) In Wikipedia. Retrieved 23-06-2014. https://en.wikipedia.org/wiki/PlayStation_3_cluster.
- [27 PlayStation3 Cluster (16-04-2010) In Wikipedia Retrieved 20-06-2014. https://en.wikipedia.org/wiki/PlayStation_3_cluster
- [28] PlayStation3 Cluster (05-03-2010)In Wikipedia. Retrieved 21-06-2014 https://en.wikipedia.org/wiki/PlayStation_3_cluster
- [29]Supercomputer (06-03-2010) In Wikipedia Retrieved 25-06-2014. https://en.wikipedia.org/wiki/Supercomputer
- [30] PlayStation3 Cluster. (04-09-2010)In Wikipedia Retrieved 25-06-2014. https://en.wikipedia.org/wiki/PlayStation_3_cluster
- [31] Black Holes and Quantum Loops (12-11-2010)In Phys.org. Retrieved 05-07-2014. http://phys.org/news/2014-02-black-holes-thought.html
- [32] Condor Supercomputer (29-11-2010) In Wpafb.af-News Retrieved 15-12-2014. http://www.wpafb.af.mil/news/story.asp?id=123231285
- [33] PlayStation3 Cluster (30-11-2010) In Wikipedia. Retrieved 28-07-2014. https://en.wikipedia.org/wiki/PlayStation_3_cluster
- [34] Folding@Home (18-03-2011) In Sie.com. Retrieved 18-07-2014. https://www.sie.com /en/corporate/release/2008/081106d.html
- [35] Folding@Home (15-05-2010) In Folding.Stanford.edu. Retrieved 28-07-2014. https://folding.stanford.edu/home/faq/
- [36]TOP500 Listing (16-11-2009)In Wikirans.net. Retrieved 09-7-2014. http://epo .wikitrans.net/PlayStation_3_cluster

  [37] PlayStation3 Cluster (15-02-2011) In Wikipedia. Retrieved 22-07-2014. https://en.wikipedia.org/wiki/PlayStation_3_cluster.
- [38] SimBioSys PlayStation3 Cluster (04-09-2011) In Wikipedia. Retrieved 22-07-2014. https://en.wikipedia.org/wiki/PlayStation_3_cluster
- [39] Bio-IT World (15-07-2012) In Bio-It worldexpo. Retrieved 06-07-2014 http://www.bio-itworldexpo.com/
- [40] Other OS.(19-09-2012) In Wikipedia. Retrieved 19-05-2013. https://en.wikipedia .org/wiki/OtherOS.
- [41] PlayStation3 Cluster (12-05-2010) In Wikipedia. Retrieved 19-04-2013. https://en.wikipedia.org/wiki/PlayStation_3_cluster

- [42] Lab´s Supercomputer PlayStation3 Cluster (19-08-2012) In Wikipedia. Retrieved 19-08-2013. https://en.wikipedia.org/wiki/PlayStation_3_cluster
- [43] Message Passing In Computer Clusters (n.d) In Wikipedia. Retrieved 17-08-2013. https://en.wikipedia.org/wiki/Message_passing_in_computer_clusters
- [44] Message Passing In Computer Cluster (n.d) In Wikipedia. Retrieved 17-08-2013. https://en.wikipedia.org/wiki/Message_passing_in_computer_clusters
- [45] Computer Cluster (n.d) In Wikipedia. Retrieved 10-05-2013. https://en.wikipedia .org/wiki/Computer_cluster
- [46] Message Passing In Computer Clusters (n.d) In Wikipedia. Retrieved 16-08-2013 https://en.wikipedia.org/wiki/Message_passing_in_computer_clusters.
- [47] Computer Cluster (n.d) In Wikipedia. Retrieved 16-08-2013. https://en.wikipedia .org/wiki/Computer_cluster
- [48] Grid Cluster Computing (n.d) In Wikipedia. Retrieved 04-08-2013 https://en.wikipedia.org/wiki/Computer_cluster
- [49] Rocks Cluster oriented Linux Distribution (n.d) In Wikipedia 09-05-2014. http://www.hpckp.org/index.php/conference/typography/68-rocks-cluster-a-cluster-oriented-linux-distribution-or-how-to-install-a-computer-cluster-in-a-day.
- [50] Computer Organisation (n.d) In Cse.hcmut.edu. Retieve 09-05-2014. http://www.cse.hcmut.edu.vn/~vtphuong/KTMT/Slides/TextBookFull.pdf.
- [51] GPU-Based Heterogenous Clusters (03-12-2010) In Johroline. Retrieved 04-05-2014. http://www.johronline.com/issue/20131212-212933.063.pdf
- [52] Fencing Computing (n.d) In Wikipedia. Retrieved 04-05-2014. https://en.wikipedia.org/wiki/Fencing_(computing).
- [53] Fencing Computing (n.d) In Wikipedia 06-05-2014. https://en.wikipedia.org/wiki /Fencing_(computing).
- [54] Computer Cluster (n.d) In Wikipedia. Retrieved 06-05-2014. https://en.wikipedia.org/wiki /Computer_cluster
- [55]Computer Cluster (05-10-2011) In Wikipedia. Retrieved 12-05-2014 https://en.wikipedia.org/wiki/Computer_cluster
- [56] A debugging standard for high-performance computing (02-04-2000) In johronline. Retrieved 12-05-2014 http://www.johronline.com/issue/20131212-212933.063.pdf
- [57] Computer Cluster (n.d) In Wikipedia 14-05-2014. https://en.wikipedia.org/wiki/ Computer_cluster
- [58] Flash Mob Computing (23-02-2004) In Wikipedia. Retrieved 14-05-2014. https://en.wikipedia.org/wiki/Flash_mob_computing.
- [59] Red Hat Cluster Suite (02-04-2012) In Wikipedia. Retrieved 14-05-2014. https://en.wikipedia.org/wiki/Red_Hat_cluster_suite.
- [60] Red Hat Cluster Suite (02-04-2012) In Wikipedia. Retrieved 15-04-2014. https://en.wikipedia.org/wiki/Red_Hat_cluster_suite.
- [61 Red Hat Cluster Suite (17-07-2009) In Wikipedia. Retrieved 14-05-2014. https://en.wikipedia.org/wiki/Red_Hat_cluster_suite.
- [62] Scalability (23-04-2012) In Cnet-News.com. Retrieved 17-05-2014. http://www.cnet.com/news/scalability-day-falls-short/.
- [63] Microsoft Cluster Server (09-06-2006) In Wikipedia. Retrieved 17-05-2014. https://en.wikipedia.org/wiki /Microsoft_Cluster_Server

- [64] Solaris Cluster (11-12-2012) In Wikipedia. Retrieved 15-07-2013. https://en.wikipedia.org/wiki/Solaris_Cluster
- [65] Solaris Cluster (20-08-2012) In Wikipedia Retrieved 20-08-2013. https://en.wikipedia.org/wiki/Solaris_Cluster
- [66] Xgrid (07-02-2004) In Wikipedia. Retrieved 18-06-2013. https://en.wikipedia .org/wiki/Xgrid.
- [67] Xgrid (15-02-2004) In Wikipedia. Retrieved 26-08-2013. http://everything.explained.today/Xgrid/
- [68] Xgrid (10-03-2008) In Wikipedia. Retrieved 18-08-2013. http://everything.explained .today/Xgrid/
- [69] Xgrid Programming Guide (31-10-2007) In Wikipedia. Retrieved 07-12-2013. https://en.wikipedia .org/wiki/Xgrid
- [70] Mac OS X Server: Xgrid (01-11-2007) In Apple.com. Retrieved 03-12-2013. http://www.apple.com/server/docs/Xgrid_Admin_v10.4.pdf
- [71]. Xgrid with Tiger Client (23-06-2005) In Macosxhints.com. Retrieved 26-07-2013. http://www.apple.com/server/docs/Xgrid_Admin_v10.4.pdf
- [72] Xgrid (23-07-2006) In Everything.explained.today. Retrieved 26-07-2013. http://everything.explained.today/Xgrid/
- [73] Xgrid (01-08-2006) In Everything.explained.today. Retrieved 03-12-2013. http://everything.explained.today/Xgrid/.

Δευτερεύουσες Αναφορές :

- Pfister, Gregory (1998). In Search of Clusters (2nd ed.). Upper Saddle River, NJ: Prentice Hall PTR. p. 36. ISBN 0-13-899709-8. https://en.wikipedia.org/wiki/ Computer_cluster,
- Readings in computer architecture by Mark Donald Hill, Norman Paul Jouppi, Gurindar Sohi 1999 ISBN 978-1-55860-539-8 page 41-48 https://en.wikipedia.org/wiki /Supercomputer
- Bornschlegl, Susanne (2012). Railway Computer 3.0: An Innovative Board Design Could Revolutionize The Market (pdf). MEN Mikro Elektronik. http://www.railway-technology .com/contractors/computer/men-mikro/
- High Availability".linuxvirtualserver.org. https://en.wikipedia.org/wiki/Load_ balancing_ (computing)

  Shuang Yu (8 May 2012). "IEEE APPROVES NEW IEEE 802.1aq™ SHORTEST PATH BRIDGING STANDARD". IEEE. https://en.wikipedia.org/wiki/IEEE_802.1aq.
- Peter Ashwood-Smith (24 Feb 2011). "Shortest Path Bridging IEEE 802.1aq Overview". Huawei. https://en.wikipedia.org/wiki/IEEE_802.1aq
- Jim Duffy (11 May 2012). "Largest Illinois healthcare system uproots Cisco to build $40M private cloud". PC Advisor. "Shortest Path Bridging will replace Spanning Tree in the Ethernet fabric." https://en.wikipedia.org/wiki/

Network_topology

- "IEEE Approves New IEEE 802.1aq Shortest Path Bridging Standard". Tech Power Up. 7 May 2012. https://en.wikipedia.org/wiki/ IEEE_802.1aq.
- High Performance Computing for Computational Science - VECPAR 2004 by Michel Daydé, Jack Dongarra 2005 ISBN 3-540-25424-2 pages 120-121. http://www.johronline.com/issue/20131212-212933.063.pdf
- Hamada T. et al. (2009) A novel multiple-walk parallel algorithm for the Barnes–Hut treecode on GPUs – towards cost effective, high performance N-body simulation. Comput. Sci. Res. Development 24:21-31. doi:10.1007/s00450-009-0089-1. https://en.wikipedia.org/wiki/DEGIMA
- "Scientific Computing on the Sony PlayStation 2". https://en.wikipedia.org /wiki/PlayStation_3_cluster
- "Terra Soft to Provide Linux for PLAYSTATION 3". https://en.wikipedia.org/wiki /PlayStation_3_cluster
- "Linux pre-installed on PS3". Terra Soft. http://www.askdavetaylor.com/yellow_dog _linux_on_sony_playstation3/
- "Linux clusters". Terra Soft. https://en.wikipedia.org/wiki/PlayStation_3_cluster
- "RapidMind and Terra Soft partner to unleash PlayStation 3 for Linux". RapidMind. 2012-6-10. https://en.wikipedia.org/wiki/Yellow_Dog_Linux
- "Engineer Creates First Academic Playstation 3 Computing Cluster". PhysOrg.com. 2012-8-9. http://phys.org/news/2007-03-academic-playstation-cluster.html
- "NC State Engineer Creates First Academic Playstation 3 Computing Cluster". College of Engineering, North Carolina State University. 2012-7-10 http://phys.org/news/2007-03-academic-playstation-cluster.html
- "Astrophysicist Replaces Supercomputer with Eight PlayStation 3s". Wired. 2010-10-17. http://archive.wired.com/techbiz/it/news/2007/10/ps3_supercomputer
- "PS3 cluster creates homemade, cheaper supercomputer" 2011-9-16 http://www.computerworld.com/article/2539437/high-performance-computing/ps3-cluster-creates-homemade--cheaper-supercomputer.html
- Highfield, Roger (2008-02-17). "Why scientists love games consoles". The Daily Telegraph (London). https://en.wikipedia.org/wiki/PlayStation_3_cluster
- The Supercomputer Goes Personal". 2010-3-6. https://en.wikipedia.org/wiki /Supercomputer
- Farrell, John (2010-11-12). "Black Holes and Quantum Loops: More Than Just a Game". 2011-5-14.
- Koff, Stephen (November 30, 2010). "Defense Department discusses new Sony PlayStation supercomputer".blog.cleveland.com.tps://en.wikipedia.org/wiki /PlayStation_3_cluster
- "Folding@Home - Client statistics by OS". Stanford University. 2011-5-15. https://folding.stanford.edu/home/faq/

ΚΕΦΑΛΑΙΟ 3

Rocks Cluster distribution

3.1 Εισαγωγή στο Rocks Cluster

Η Διανομή Rocks Cluster (που αρχικά ονομαζόταν NPACI Rocks) είναι μια διανομή Linux που προορίζεται για High Performance computing clusters. Άρχισε από την Εθνική Σύμπραξη για Σύνθετους υπολογισμούς (Advanced Computational Instructure) και το San Diego Supercomputer Center (SDSC) το 2000[1] και αρχικά χρηματοδοτήθηκε εν μέρει από μια επιχορήγηση της NSF (Ntional Science Foundation) (2000-2007)[2] , αλλά χρηματοδοτήθηκε και από το follow-up NSF μέσω επιχορήγησης το 2011.[3] Το Rocks βασίστηκε αρχικά στην διανομή Red Hat Linux, ωστόσο, σύγχρονες εκδόσεις των Rocks βασίστηκαν σε CentOS, με ένα τροποποιημένο πρόγραμμα εγκατάστασης το Anaconda που απλοποιεί τη μαζική εγκατάσταση σε πολλούς υπολογιστές. Το Rocks περιλαμβάνει πολλά εργαλεία (όπως το MPI) που δεν αποτελούν μέρος του CentOS, αλλά αποτελούν αναπόσπαστα συστατικά που μετατρέπουν μια ομάδα υπολογιστών σε Cluster.

Οι εγκαταστάσεις μπορεί να προσαρμοστούν με πρόσθετα πακέτα λογισμικού κατά τον χρόνο εγκατάστασης (install-time) με τη χρήση ειδικών CDs παρεχόμενο από το χρήστη (που ονομάζεται "Roll CDs"). Το «Roll επιτρέπει την επέκταση του συστήματος με την απρόσκοπτη ενσωμάτωση, αυτόματους μηχανισμούς διαχείρισης και τις συσκευασίες που χρησιμοποιούνται από τη βάση του λογισμικού, απλοποιώντας σημαντικά την εγκατάσταση και διαμόρφωση του μεγάλου αριθμού των υπολογιστών.[4] Πάνω από μια ντουζίνα Rolls έχουν δημιουργηθεί, συμπεριλαμβανομένης του SGE roll, το Condor roll, το Lustre Roll, το Java roll, και το Ganlia roll. Μέχρι τον Οκτώβριο του 2010, το Rocks χρησιμοποιήθηκε από ακαδημαϊκούς, κυβερνητικούς και εμπορικούς οργανισμούς, όπου απασχολούσε με 1.376 Cluster, σε όλες τις ηπείρους εκτός από την Ανταρκτική.[5] Το μεγαλύτερο ακαδημαϊκό Cluster που έχει καταχωρηθεί, με 8632 επεξεργαστές, είναι GridKa, που λειτουργεί από το Τεχνολογικό Ινστιτούτο της Καρλσρούης στην Γερμανία (Karlsruhe Institute of Technology).

Υπάρχουν επίσης μια σειρά από Cluster που αποτελούνται το πολύ από δέκα επεξεργαστές. Αντιπροσωπεύουν τα πρώτα στάδια της κατασκευής των μεγαλύτερων συστημάτων cluster.

Αυτή η εύκολη επεκτασιμότητα ήταν ένας σημαντικός παράγοντας για την ανάπτυξη των Rocks, τόσο για τους συμμετέχοντες ερευνητές, όσο και για την NSF.

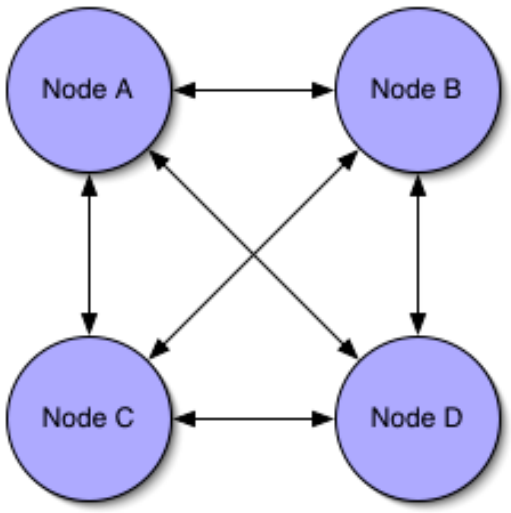
3.2 Επίπεδο Επικοινωνίας Rocks Cluster (Comunication Layer)

- **Γενικά**
  - Υποστήριξη Παραλληλισμού

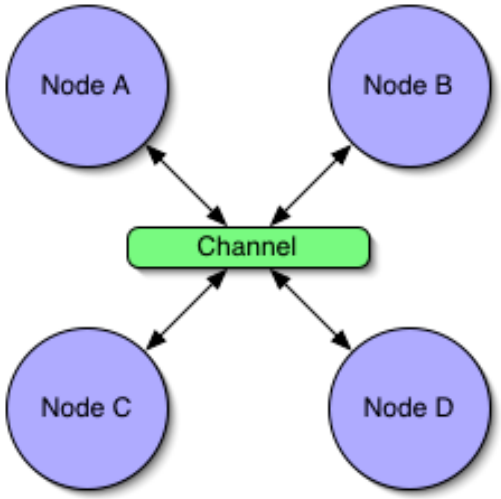
- **Sockets (Υποδοχές)**
  - Μοντέλο Πελάτη-Εξυπηρετητή (Client Server model)
  - Point-to-Point επικοινωνία

- **MPI- Message Passing Interface**
  - Message Passing
  - Στατικό μοντέλο των συμμετεχόντων (Στατικό Μοντέλο)

- **PVM- Parallel Virtual Machines**
  - Message Passing
  - Για ετερογενείς αρχιτεκτονικές
  - Έλεγχος πόρων και Ανοχή σφαλμάτων



Σχήμα: 3.1 Sockets, Point-to-Point. N machines = (n^2 –n)/2 συνδέσεις. 1, 3, 6, 10, 15,…



Σχήμα 3.2 MPI/PVM. Διαμοιρασμός εικονικού καναλιού (Virtual Channel). Δυνατότητα εφαρμογής sockets, καθώς και εύκολος προγραμματισμός.

3.2.1 Network Socket

Ένα Network Socket είναι ένα τελικό σημείο μιας ροής επικοινωνίας μεταξύ διεργασιών σε ένα δίκτυο υπολογιστών. Σήμερα, η περισσότερη επικοινωνία μεταξύ υπολογιστών βασίζεται στο Internet Protocol. Ως εκ τούτου, τα περισσότερα network sockets είναι Internet sockets.

Ένα socket API είναι μια διεπαφή προγραμματισμού εφαρμογών (API), που συνήθως παρέχεται από το λειτουργικό σύστημα, το οποίο επιτρέπει στην εφαρμογή προγραμμάτων, τον έλεγχο και τη χρήση των network sockets.Τα Internet socket APIs συνήθως βασίζονται στο πρότυπο Berkeley sockets.

Ένα socket address είναι ο συνδυασμός μιας διεύθυνσης IP και ο αριθμό θύρας (port), μοιάζει πολύ με το ένα άκρο της τηλεφωνικής σύνδεσης που είναι ο συνδυασμός ενός αριθμού τηλεφώνου και μια συγκεκριμένη επέκταση. Με βάση αυτή τη διεύθυνση, τα internet sockets παραδίδουν τα εισερχόμενα πακέτα δεδομένων με την κατάλληλη διαδικασία υποβολής αιτήσεων ή νημάτων (threads).

Ένα internet socket χαρακτηρίζεται από έναν μοναδικό συνδυασμό από τα ακόλουθα:

- Τοπική διεύθυνση socket : Τοπική IP διεύθυνση και αριθμός θύρας.

- Απομακρυσμένη διεύθυνση socket: Μόνο για προκαθορισμένα TCP sockets. Όπως αναφέρθηκε στο τμήμα client-server, αυτό είναι αναγκαίο, δεδομένου ότι ένας διακομιστής TCP μπορεί να εξυπηρετήσει πολλούς πελάτες ταυτόχρονα. Ο διακομιστής δημιουργεί μια

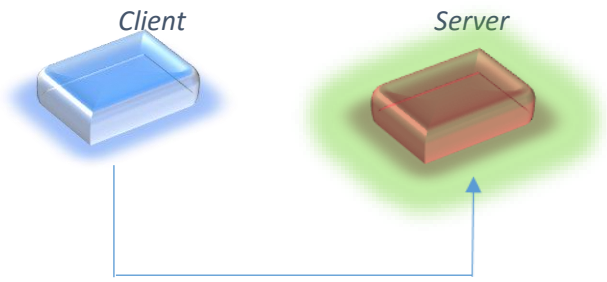
υποδοχή (socket) για κάθε πελάτη, και αυτές οι υποδοχές μοιράζονται την ίδια τοπική διεύθυνση υποδοχή από την άποψη του διακομιστή TCP.

- Πρωτόκολλο: Σε ένα πρωτόκολλο μεταφοράς (π.χ., TCP, UDP, raw IP, ή άλλα). Η TCP θύρα (port) 53 και UDP θύρα 53 είναι, κατά συνέπεια, διαφορετικές, ξεχωριστές υποδοχές (sockets).

  Εντός του λειτουργικού συστήματος και της εφαρμογής που δημιουργείτε μια υποδοχή (socket), το socket αναφέρεται σε μια μοναδική ακέραιη τιμή (integer) που ονομάζεται socket descriptor. Το λειτουργικό σύστημα προωθεί το ωφέλιμο φορτίο των εισερχόμενων πακέτων IP με την αντίστοιχη εφαρμογή με την εξαγωγή της διεύθυνσης υποδοχής πληροφορίας από τις IP και επικεφαλίδα μεταφοράς πρωτόκολλου και διαχωρισμό της επικεφαλίδας από τα δεδομένα της εφαρμογής.

  Λειτουργία Socket:

  - Άνοιγμα ενός τελικού σημείου (end point).
  - Καθορισμός διεύθυνσης IP και θύρας (port).
  - Αποστολή και παραλαβή μηνυμάτων.
    - Εάν χρησιμοποιεί TCP, τότε μόνο μηνύματα point-to-point.
    - Εάν χρησιμοποιεί UDP, τότε υπάρχει δυνατότητα επιλογής μεταξύ point-to-point ή multicast (broadcast).
  - Κλείσιμο Σύνδεσης.



Σχήμα: 3.3 Μοντέλο Πελάτης-Εξυπηρετητής

3.2.2 Διαφορές TCP και UDP sockets

- **TCP**
  - Αξιόπιστο, αλλά byte oriented (προσανατολισμένο)
  - Χρειάζεται να γράψει κώδικα. Ώστε να στείλει και να λάβει πακέτα (στο επίπεδο Εφαρμογών)

- **UDP**
  - Αναξιόπιστο
  - Χρειάζεται να γράψει κώδικα για να στείλει αξιόπιστα πακέτα.

3.3 Message Passing Interface (MPI)

- Message Passing Interface.

- De facto πρότυπο για την ανταλλαγή μηνυμάτων.
  - Τρέχει σε πολλές αρχιτεκτονικές CPU και σε πολλά επικοινωνιακά υποστρώματα.

- . Υπάρχουν (και υπήρχαν) πολλές καλές βιβλιοθήκες μηνυμάτων.
  - Αλλά, η MPI είναι η πιο διαδεδομένη.
  - Ανέπτυξε ένα πρακτικό, φορητό, αποτελεσματικό και ευέλικτο πρότυπο.
  - Σε εξέλιξη από το 1992

Το Message Passing Interface (MPI) είναι ένα τυποποιημένο και φορητό message passing σύστημα που σχεδιάστηκε από μια ομάδα ερευνητών από τον ακαδημαϊκό χώρο και τη βιομηχανία για να λειτουργεί σε μια ευρύ ποικιλία των παράλληλων υπολογιστών. Το πρότυπο ορίζει το συντακτικό και τη σημασιολογία ενός πυρήνα της βιβλιοθήκης με τις ρουτίνες, χρήσιμα σε ένα ευρύ φάσμα χρηστών γραφής φορητών προγραμμάτων ανταλλαγής μηνυμάτων σε διάφορες γλώσσες προγραμματισμού ηλεκτρονικών υπολογιστών, όπως η Fortran, C, C ++ και Java. Υπάρχουν πολλές καλά δοκιμασμένες και αποτελεσματικές υλοποιήσεις του MPI, συμπεριλαμβανομένων και ορισμένων που είναι δωρεάν ή στον δημόσιο τομέα. Αυτό προώθησε την ανάπτυξη μιας παράλληλης βιομηχανίας λογισμικού, και ενθάρρυνε να υπάρχει ανάπτυξη των φορητών και μεγάλης κλίμακας παράλληλων εφαρμογών.

3.3.1 Επισκόπηση του MPI

Το MPI είναι μια γλώσσα ανεξάρτητη από το πρωτόκολλο επικοινωνίας που χρησιμοποιείται για τον προγραμματισμό παράλληλων υπολογιστών. Τόσο point-to-point όσο και συλλογικής επικοινωνίας υποστηρίζονται. Το MPI είναι μία εφαρμογή message passing, σε συνδυασμό με το πρωτόκολλο και σημασιολογικές προδιαγραφές για το πώς τα χαρακτηριστικά του πρέπει να συμπεριφέρονται σε κάθε εφαρμογή.[6] Τα πλεονεκτήματα του MPI είναι υψηλή απόδοση, επεκτασιμότητα και φορητότητα. Το MPI παραμένει το κυρίαρχο μοντέλο που χρησιμοποιείται στα πληροφοριακά συστήματα υψηλής απόδοσης (High Performance Computing) σήμερα.[7]

Το MPI δεν αποκλείεται από κανένα σημαντικό οργανισμό τυποποίησης, παρ' όλα αυτά, έχει γίνει ένα de facto πρότυπο για την επικοινωνία μεταξύ διεργασιών του μοντέλου όπου ένα παράλληλο πρόγραμμα τρέχει σε ένα κατανεμημένο σύστημα μνήμης. Στην πραγματικότητα οι υπερυπολογιστές κατανεμημένης μνήμης: όπως είναι τα computer cluster τρέχουν συχνά τέτοια προγράμματα. Το κύριο MPI-1 μοντέλο δεν έχει κοινή αντίληψη της μνήμης, και το MPI-2 έχει μόνο μια περιορισμένη έννοια της κατανεμημένης μνήμης. Παρ 'όλα αυτά, τα προγράμματα MPI συνήθως τρέχουν σε κατανεμημένους υπολογιστές μνήμης. Σχεδιάζοντας προγράμματα γύρω από το μοντέλο MPI (Σε αντίθεση με τα συνηθισμένα μοντέλα μνήμης) έχουν πλεονεκτήματα σε σχέση με NUMA αρχιτεκτονικές μνήμης, όσο το MPI υποστηρίζει την τοπικότητα της μνήμης.

Επίσης το MPI ανήκει στο στρώμα 5 και υψηλότερα του μοντέλου αναφοράς OSI, οι εφαρμογές μπορεί να καλύπτουν περισσότερες στρώσεις, με υποδοχές-sockets και το πρωτόκολλο ελέγχου μετάδοσης (TCP) που χρησιμοποιείται στο στρώμα μεταφοράς.

Οι περισσότερες υλοποιήσεις MPI αποτελούνται από ένα συγκεκριμένο σύνολο από ρουτίνες (δηλαδή, ένα API) άμεσα καλούμενες από C, C ++, Fortran και οποιαδήποτε γλώσσα σε θέση να διασυνδέονται με τις βιβλιοθήκες, συμπεριλαμβανομένων των C#, Java ή Python. Τα πλεονεκτήματα του MPI σε σχέση με παλαιότερες βιβλιοθήκες message passing, είναι η φορητότητα (γιατί το MPI έχει εφαρμοστεί σχεδόν σε κάθε κατανεμημένη αρχιτεκτονική μνήμης) και η ταχύτητα (επειδή κάθε εφαρμογή έχει κατ' αρχήν βελτιστοποιηθεί για το hardware στο οποίο τρέχει). Το MPI χρησιμοποιεί προδιαγραφές ανεξαρτήτου γλώσσας προγραμματισμού (Language Independent Specifications-LIS) για κλήσεις και διασύνδεση (binding) με κάποια γλώσσα. Το πρώτο πρότυπο MPI ορίστηκε από ANSI C και Fortran-77 «δεμένα» μαζί με το LIS. Το σχέδιο παρουσιάστηκε στο Supercomputing 1994.[8] Περίπου

128 λειτουργίες αποτελούν το πρότυπο MPI-1.3 που κυκλοφόρησε ως το οριστικό τέλος της σειράς MPI-1 το 2008.[9]

Προς το παρόν, το πρότυπο αυτό έχει αρκετές εκδοχές: έκδοση 1.3 (κοινώς συντομογραφία MPI-1), η οποία τονίζει το μήνυμα που περνά και έχει ένα στατικό περιβάλλον χρόνου εκτέλεσης, MPI-2.2 (MPI-2), το οποίο περιλαμβάνει νέα χαρακτηριστικά, όπως η παράλληλη είσοδος/έξοδος(I/O), λειτουργίες δυναμικής διαχείρισης και διεργασίες απομακρυσμένης μνήμης και MPI-3.0 (MPI-3),[10] το οποίο περιλαμβάνει επεκτάσεις των συλλογικών εργασιών με μη δεσμευτικές εκδόσεις και επεκτάσεις στις μονόπλευρες διεργασίες.[11] Το LIS MPI-2, διευκρινίζει πάνω από 500 λειτουργίες και παρέχει γλώσσες «δέσμευσης» για ANSI C, ANSI C ++, και ANSI Fortran(Fortran90). Η Διαλειτουργικότητα Αντικείμενων (Object Interoperability) προστέθηκε επίσης για να επιτρέψει την ευκολότερη προώθηση μηνυμάτων μεταξύ προγραμμάτων γραμμένα σε διαφορετικές γλώσσες προγραμματισμού.

Το MPI-2 είναι ως επί το πλείστων ένα υπερσύνολο του MPI-1, αν και ορισμένες λειτουργίες έχουν απενεργοποιηθεί. Τα MPI-1.3 προγράμματα εξακολουθούν να λειτουργούν σύμφωνα με υλοποιήσεις του MPI που συμμορφώνεται με το πρότυπο MPI-2.

Το MPI-3 περιλαμβάνει τη Fortran 2008 γλώσσα binding, ενώ αφαιρεί απενεργοποιημένες C ++ «δεσμεύσεις» καθώς και πολλές απενεργοποιημένες ρουτίνες MPI και αντικείμενα.

Το MPI είναι συχνά συγκρίσιμο με το Parallel Virtual Machine(PVM), το οποίο είναι ένα δημοφιλές κατανεμημένο περιβάλλον και το σύστημα του, το message passing αναπτύχθηκε το 1989, καθώς ήταν ένα από τα συστήματα που προκαλέσανε την ανάγκη για τυποποίηση μοντέλου παράλληλου message passing. Μοντέλα προγραμματισμού κοινής μνήμης (όπως Pthreads και OpenMP) και message passing προγραμματισμού (MPI/PVM) μπορούν να θεωρηθούν ως συμπληρωματικές προσεγγίσεις προγραμματισμού, και μπορεί περιστασιακά να τις δούμε μαζί σε εφαρμογές, π.χ. σε servers με πολλούς μεγάλους κόμβους και κοινόχρηστη μνήμη.

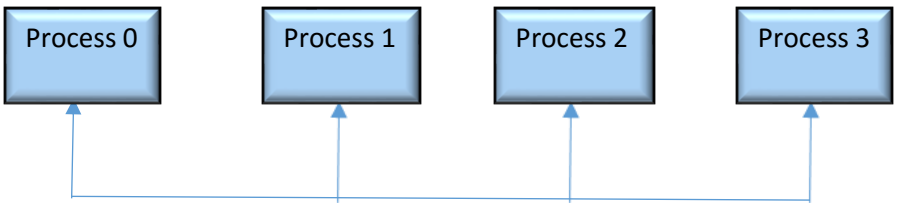
3.3.2 Λειτουργία του MPI

Η διασύνδεση MPI έχει ως στόχο να παρέχει την απαραίτητη εικονική τοπολογία, συγχρονισμό και τη λειτουργικότητα της επικοινωνίας μεταξύ ενός συνόλου από διαδικασίες (που έχουν αντιστοιχιστεί σε κόμβους-nodes / servers / υπολογιστικές μονάδες) σε μια γλώσσα

ανεξάρτητη, με συγκεκριμένη σύνταξη γλώσσας, συν μερικά ειδικά χαρακτηριστικά γλώσσας. Τα προγράμματα MPI λειτουργούν πάντα με διαδικασίες, αλλά οι προγραμματιστές συνήθως αναφέρονται στις διαδικασίες ως επεξεργαστές. Συνήθως για μέγιστη απόδοση, κάθε CPU (ή πυρήνα σε ένα multi-core μηχάνημα) θα πρέπει να ανατεθεί μόνο μια ενιαία διαδικασία. Αυτή η ανάθεση γίνεται κατά το χρόνο εκτέλεσης μέσω του παράγοντα που ξεκινά το πρόγραμμα MPI, συνήθως ονομάζεται mpirun ή mpiexec.

Βασικές προϋποθέσεις:

- Σαφώς τα δεδομένα πρέπει να μετακινούνται σαν socket (υποδοχές), αλλά πρέπει να περνάνε την διαδικασία virtualize τα τελικά σημεία (endpoints) .
    - Τα τελικά σημεία (endpoints) διευθυνσιοδοτούνται με έναν ακέραιο 0, 1, …,n.
- Πρωταρχικά θα πρέπει να υποστηρίζει Point-to-Point και Broadcast.



Σχήμα: 3.4 Καταμερισμός διεργασιών σε Point-to-Point

Η MPI βιβλιοθήκη περιλαμβάνει συναρτήσεις, αλλά δεν περιορίζεται μόνο σε αυτές, περιλαμβάνει επίσης point-to-point συνδέσεις τύπου αποστολή/λήψης, επιλέγοντας ανάμεσα σε ένα καρτεσιανό δέντρο ή γράφημα που μοιάζει λογική διεργασία τοπολογίας, επίσης η διαδικασία ανταλλαγής δεδομένων μεταξύ των ζευγαριών (αποστολή/λήψη εργασιών), συνδυάζοντας μέρος των αποτελεσμάτων των υπολογισμών (συγκεντρωθούν και να περιορίσουν τη λειτουργία), το συγχρονισμό των κόμβων (λειτουργία φραγμού), καθώς και την απόκτηση πληροφοριών σχετικά με το δίκτυο, όπως ο αριθμός των διεργασιών στη σύνοδο των υπολογιστών, την τρέχουσα ταυτότητα επεξεργαστή, ότι η διαδικασία έχει αντιστοιχηθεί σε γειτονικές διεργασίες πρόσβασης σε μία λογική τοπολογία και ούτω καθεξής. Point-to-point ενέργειες έρχονται σε σύγχρονη, ασύγχρονη, buffered και έτοιμες φόρμες, για να επιτρέψει και τα δύο, δηλαδή σημασιολογικά ισχυρότερα και ασθενέστερα για τα θέματα συγχρονισμού ενός

«ραντεβού»-αποστολής. Πολλές εκκρεμείς πράξεις είναι δυνατόν να τοποθετηθούν σε ασύγχρονη λειτουργία, στις περισσότερες εφαρμογές.

Το MPI-1 και το MPI-2 και τα δύο επιτρέπουν επικαλύπτομεντες εφαρμογές επικοινωνίας και υπολογισμού, αλλά η πράξη και η θεωρία διαφέρει. Το MPI ορίζει επίσης το νήματα ασφαλείας διεπαφών, οι οποίες έχουν συνοχή και σύζευξης στρατηγικές που βοηθούν στην αποφυγή κρυμμένων καταστάσεων στο εσωτερικό της διεπαφής. Είναι σχετικά εύκολο να γράψει πολυνηματικό Point-to-Point κώδικα MPI, και μερικές υλοποιήσεις υποστηρίζουν τον εν λόγω κώδικα. Πολυνηματική συλλογική επικοινωνία επιτυγχάνεται καλύτερα με πολλαπλά αντίγραφα των Communicators, όπως περιγράφεται παρακάτω.

3.3.3 Communicator

Ο Communicator συνδέει αντικείμενα και συνθέτει ομάδες διεργασιών στη σύνοδο MPI. Κάθε πληροφοριοδότης δίνει σε κάθε διαδικασία που περιλαμβάνεται ένα ανεξάρτητο αναγνωριστικό και οργανώνει τις διαδικασίες που περιέχει σε μία τακτοποιημένη τοπολογία. Το MPI έχει επίσης ρητές ομάδες, αλλά αυτές είναι κυρίως καλές για την οργάνωση και την αναδιοργάνωση των ομάδων των διαδικασιών πριν από ένα άλλος communicator φτιαχτεί. Το MPI καταλαβαίνει μόνο μία ομάδα ενδοεπικοινωνιακών (intracommunicator) διεργασιών και αμφίδρομης επικοινωνίας intercommunicator. Στο MPI-1, χωριστές ομάδες διεργασιών είναι πιο διαδεδομένες. Οι διμερείς πράξεις συνήθως εμφανίζονται στο MPI-2, όπου περιλαμβάνουν συλλογική επικοινωνία και δυναμική διαχείριση κατά τη διαχείριση.

3.3.4 Προκλήσεις με το MPI

- Εάν ένας κόμβος (node) αποτύχει, δεν υπάρχει εύκολος τρόπος για να τον αναμορφώσουν και να δρομολογήσουν παρακάμπτοντας το πρόβλημα.
    - Βασικά το πρόγραμμα σταματάει να λειτουργεί.
- Δυσκολίες στην διαχείριση κατά την επέκταση του cluster.
    - Δίκτυα X Μεταγλωττιστές = MPI binaries
    - Αποτέλεσμα είναι διάφορες εκδόσεις του MPI προσαρμοσμένες για Cluster.

3.4 Network File System (NFS)

Σύστημα Αρχείων Δικτύου (Network File System-NFS) είναι ένα κατανεμημένο πρωτόκολλο συστήματος αρχείων που αρχικά αναπτύχθηκε από την Sun Microsystems το 1984,[12] και επιτρέπει στο χρήστη ενός υπολογιστή-πελάτη (client) να αποκτήσει πρόσβαση σε αρχεία. Το NFS, όπως και πολλά άλλα πρωτόκολλα, στηρίζεται στο σύστημα πληροφορικής ανοικτού δικτύου Remote Procedure Call (RPC ONC-Open Network Computing). Το σύστημα αρχείων δικτύου είναι ένα ανοιχτό πρότυπο που καθορίζεται στα RFC (Request For Comments), που επιτρέπει σε οποιονδήποτε να εφαρμόσει το πρωτόκολλο.

3.4.1 Πλατφόρμες χρήσεις

Το NFS συχνά χρησιμοποιείται με Unix συστήματα (όπως το Solaris, AIX και HP-UX) και Unix-like λειτουργικά συστήματα (όπως το Linux και FreeBSD). Είναι επίσης διαθέσιμο σε λειτουργικά συστήματα όπως το κλασικό Mac OS, OpenVMS, IBM i, ορισμένες εκδόσεις των Microsoft Windows, και Novell NetWare. Εναλλακτικά πρωτόκολλα πρόσβασης απομακρυσμένων αρχείων περιλαμβάνει το Server Message Block (SMB, επίσης γνωστή ως CIFS), η Apple Filing Protocol (AFP), NetWare Core Protocol (NCP), και το File Server file system OS / 400 (QFileSvr.400).

Τα SMB και NetWare Core Protocol (NCP) εμφανίζονται πιο συχνά από ό, τι το NFS για συστήματα που εκτελούν τα Microsoft Windows, επίσης το AFP εμφανίζεται πιο συχνά από ό, τι NFS σε συστήματα Macintosh, και το QFileSvr.400 ήταν το ποιο διαδεδομένο για IBM i συστήματα. Το Haiku πρόσθεσε πρόσφατα την υποστήριξη NFSv4 ως μέρος της Google Summer of Code project.

3.4.2 Βασικά χαρακτηριστικά του NFS για Rocks Cluster

- Ο λογαριασμός του χρήστη εξυπηρετείται από το NFS.
  - Δουλεύει για μικρά cluster (<=128 nodes).
  - Δεν λειτουργεί για μεγάλα cluster (>1024 nodes).
  - Network Attached Storage (NAS) παρέχει πρόσβαση σε δεδομένα ενός δικτύου σε ετερογενή συστήματα σαν file server και όχι μόνο.
    - Το Rocks χρησιμοποιεί το Fronted υπολογιστή (¨Server¨ του cluster) για να εξυπηρετεί το NFS.
    - Συμμετοχή του NAS σε πολλά cluster.

- Μερικές εφαρμογές δεν εξυπηρετούνται από το NFS.
    - /usr/local/ δεν υπάρχει
    - Όλο το λογισμικό αποθηκεύεται τοπικά από το RPM (RedHat Package Manager).

3.5 Simple Network Management Protocol (SNMP)

Το Simple Network Management Protocol (SNMP) είναι μέρος της σουίτας προτοκόλλων Internet (IP - Internet Protocol), όπως έχει ορισθεί από το Internet Engineering Task Force (IETF). Χρησιμοποιείται στα συστήματα διαχείρισης δικτύων, στη διαχείριση και παρακολούθηση δικτυακών συσκευών που απαιτούν παρέμβαση του διαχειριστή δικτύου. Αποτελείται από μια ομάδα προτύπων για τη διαχείριση δικτύου και περιλαμβάνει ένα πρωτόκολλο επιπέδου εφαρμογών (application layer), ένα σχήμα βάσης δεδομένων και μια ομάδα από σύνολα δεδομένων.[13]

3.5.1 Σύνοψη και βασικοί όροι SNMP

Σε μια τυπική χρήση του πρωτοκόλλου SNMP, υπάρχει ένας αριθμός συστημάτων υπό διαχείριση καθώς και ένα ή περισσότερα συστήματα διαχείρισης. Το λογισμικό που "τρέχει" σε κάθε δικτυακή υπό διαχείριση συσκευή ή σύστημα ονομάζεται agent και αναφέρει μέσω του πρωτοκόλλου SNMP στα συστήματα διαχείρισης.

Στη βασική του μορφή, το SNMP προσφέρει στα συστήματα διαχείρισης, δεδομένα διαχείρισης ως μεταβλητές, όπως πχ."free memory", "system name", "number of running processes", "default route", κλπ. Ταυτόχρονα, επιτρέπει ενέργειες όπως η εφαρμογή νέας ή η αλλαγή της υπάρχουσας παραμετροποίησης της δικτυακής διάταξης.

Οι μεταβλητές που ελέγχονται από το SNMP οργανώνονται σε ιεραρχικές δομές, οι οποίες μαζί με τα μεταδεδομένα (όπως ο τύπος και η περιγραφή των μεταβλητών περιγράφονται από τα MIBs (Management Information Bases).

- Ενεργοποιείται σε όλους τους υπολογιστικούς κόμβους (compute nodes).

- Πολύ καλό για point-to-point χρήση.
    - Καλό για υψηλή ποιότητα σε ένα ενιαίο end-point.
    - Δεν υπάρχει κλιμάκωση σε ένα πήρες cluster ευρείας χρήσης.

- Υποστηρίζει Linux MIB (Management Information Bases).

- Uptime, Load, Στατιστικά διαδικτύου.
- Εγκατάσταση Λογισμικού.
- «Τρέξιμο» Διεργασιών.

3.6 Syslog

Το Syslog είναι ένα πρότυπο για την ιδιοποίηση μέσο μηνύματος για το logging των υπολογιστών. Επιτρέπει το διαχωρισμό του λογισμικού που παράγει μηνύματα από το σύστημα που τα αποθηκεύει και το λογισμικό που αναφέρει τις αναλύσεις τους.

Το Syslog μπορεί να χρησιμοποιηθεί για τη διαχείριση του συστήματος του υπολογιστή και ελέγχου ασφαλείας, καθώς και γενικευμένα ενημερωτικά μηνύματα, την ανάλυση και τον εντοπισμό σφαλμάτων. Η προσπάθεια αυτή υποστηρίζεται από μια μεγάλη ποικιλία από συσκευές (όπως εκτυπωτές και routers) και δέκτες σε πολλαπλές πλατφόρμες. Χάρη σε αυτό, το syslog μπορεί να χρησιμοποιηθεί για την ενσωμάτωση δεδομένων καταγραφής από πολλούς διαφορετικούς τύπους συστημάτων σε ένα κεντρικό αποθηκευτικό χώρο.

Τα μηνύματα που επισημαίνονται με τον κωδικό εγκατάστασης (μερικά από αυτά είναι: auth, authPriv, daemon, cron, ftp, lpr, kern, mail, news, syslog,user, uucp, local0 ... local7) που υποδεικνύει τον τύπο του λογισμικού που δημιουργούνται τα μηνύματα, και τους αποδίδεται βαρύτητα (Όπως είναι: Emergency (εκτάκτου ανάγκης), Alert (συναγερμός), Critical (κρίσιμο), Error (λάθος), Warning (προειδοποίηση), Notice (επισήμανση), Info (πληροφορίες) and Debug (αποσφαλμάτωση)).

Οι εφαρμογές είναι διαθέσιμες για πολλά λειτουργικά συστήματα. Ειδική διαμόρφωση μπορεί να επιτρέψει την κατεύθυνση των μηνυμάτων σε διάφορες συσκευές (console), τα αρχεία (/var/log/) ή απομακρυσμένους syslog servers. Οι περισσότερες εφαρμογές παρέχουν επίσης ένα βοηθητικό πρόγραμμα γραμμής εντολών (comand line), που συχνά αποκαλείται logger, που μπορεί να στείλει μηνύματα προς το syslog. Μερικές εφαρμογές επιτρέπουν το φιλτράρισμα και την εμφάνιση των syslog μνημάτων.

3.6.1 Σύνοψη και βασικά χαρακτηριστικά Syslog

- Native Unix σύστημα καταγραφής logger.
  - Καταγραφή γεγονότων στον τοπικό δίσκο.

    • /var/log/message

• Επαναφορά των logs καθημερινά, καθώς το ιστορικό των δεδομένων χάνεται.

- Προώθηση όλων των μηνυμάτων στον Fronted σύστημα του cluster.

- Επέκταση
  - Μπορεί να προσθέσει επιπλέον loghosts.
  - Μπορεί να περιορίσει την «φλυαρία» των loggers.

- Χρήσεις
  - Προβλέψεις για τυχών αποτυχίες σε Υλικό (Hardware) και Λογισμικό (Software).
  - Post Mortem («νεκροψία») στους κόμβους όπου σταμάτησαν να λειτουργούν λόγω κάποιου προβλήματος ή μιας σειράς προβλημάτων.
  - Αποσφαλμάτωση του συστήματος στο ξεκίνημα (startup).

3.7 Ganglia

Το Ganglia είναι ένα επεκτάσιμο εργαλείο σε κατανεμημένα συστήματα (distributed), για υψηλής απόδοσης υπολογιστικά συστήματα όπως είναι τα cluster και τα grids συστήματα. Επιτρέπει στον χρήστη να βλέπει εξ αποστάσεως ζωντανά ή ιστορικά στατιστικά στοιχεία (όπως η μέση τιμή φορτίου της CPU ή του δικτύου) για όλες τις μηχανές που παρακολουθούνται.

Το Ganglia είναι βασισμένο σε μια ιεραρχική σχεδίαση και απευθύνεται στις κοινότητες των clusters. Στηρίζεται σε multicast πρωτόκολλο μια listen/announced (άκουσε/ανακοίνωσε) όπου βασίζεται στην παρακολούθηση (monitoring) της κατάστασης εντός των cluster και χρησιμοποιεί ένα δέντρο με σημείου-προς-σημείο (point-to-point) συνδέσεις μεταξύ των αντιπροσωπευτικών κόμβων του cluster, για να συνασπίσει το clusters και να συγκεντρώνει την κατάσταση τους. Αξιοποιεί διαδεδομένες τεχνολογίες όπως XML για την αναπαράσταση των δεδομένων, XDR (eXternal Data Representaion) για την μεταφορά αρχείων σε διαφορετικού τύπου συστήματα , Portable Data Transport, και RRDtool για την αποθήκευση δεδομένων και την απεικόνιση. Χρησιμοποιεί προσεκτικά σχεδιασμένες δομές δεδομένων και αλγορίθμους για την επίτευξη πολύ χαμηλά ανά-κόμβο κόστος χρήσης και υψηλό συγχρονισμό. Η εφαρμογή είναι πολύ ισχυρή, έχει μεταφερθεί σε ένα εκτεταμένο σύνολο των

λειτουργικών συστημάτων και αρχιτεκτονικές επεξεργαστή, και είναι σήμερα σε χρήση για πάνω από 500 clusters σε όλο τον κόσμο. Έχει χρησιμοποιηθεί για τη σύνδεση των clusters σε πανεπιστημιουπόλεις, σε όλο τον κόσμο και μπορεί να επεκταθεί για να χειριστεί clusters ακόμα και με 2000 κόμβους. [14]

Το σύστημα Ganlia περιλαμβάνει δύο μοναδικά daemons, ένα PHP με βάση το web front-end, και μερικά άλλα μικρότερα προγράμματα-utilities. Όπως περιγράφονται παρακάτω.

3.7.1 Ganglia Monitoring Daemon (gmond)

Gmond είναι ένα multi-threaded daemon που τρέχει σε κάθε κόμβο του cluster που θέλουμε να παρακολουθήσουμε. Δεν απαιτεί εγκατάσταση έχοντας ένα κοινό σύστημα αρχείων NFS ή μια βάση δεδομένων back-end, εγκαθιστώντας ειδικούς λογαριασμούς ή τη διατήρηση των αρχείων διαμόρφωσης.

Το Gmond έχει τέσσερις βασικές αρμοδιότητες:

1. Παρακολουθούν τις αλλαγές στο κατάσταση υποδοχής (host state).
2. Ανακοινώνει τις διάφορες αλλαγές.
3. Να «ακούει» την κατάσταση όλων των άλλων ganglia κόμβων (nodes) μέσω unicast ή multicast κανάλι.
4. Απαντά στα αιτήματα για μια XML περιγραφή της κατάστασης cluster.

Κάθε gmond μεταδίδει τις πληροφορίες με δύο διαφορετικούς τρόπους.

- Μονή εκπομπή ή Πολλαπλή κατάσταση υποδοχής στην εξωτερική μορφή εκπροσώπησης δεδομένων (eXternal Data Representation-XDR) χρησιμοποιώντας UDP μηνύματα.
- Η αποστολή XML γίνεται μέσω μιας σύνδεσης TCP.

3.7.2 Ganglia Meta Daemon (gmetad)

Η ενότητα-ομοσπονδία στο Ganglia, επιτυγχάνεται χρησιμοποιώντας ένα δέντρο με σημείου-προς-σημείο συνδέσεις μεταξύ των αντιπροσωπευτικών κόμβων του cluster, για να συγκεντρώσει την κατάσταση των πολλαπλών cluster. Σε κάθε κόμβο (node) στο δέντρο, ένα Ganglia Meta Daemon (gmetad) περιοδικά δημοσκοπεί μια συλλογή των πηγέων δεδομένων, αναλύει συνολικά τα XML δεδομένα, αποθηκεύει όλα τα αριθμητικά, επίσης κάνει μετρήσεις σε βάσεις δεδομένων round-robin και εξάγει το σύνολο των δεδομένων XML πάνω από μία υποδοχή TCP για τους πελάτες (clients). Οι πηγές των δεδομένων μπορεί να είναι είτε gmond

daemin, που εκπροσωπούν συγκεκριμένα cluster ή άλλοι daemon gmetad, που εκπροσωπούν σύνολα από cluster. Οι πηγές δεδομένων χρησιμοποιούν τις διευθύνσεις IP πηγής για τον έλεγχο της πρόσβασης και μπορεί να προσδιορίζονται χρησιμοποιώντας πολλαπλές διευθύνσεις IP για ανακατεύθυνση. Η τελευταία αυτή δυνατότητα είναι φυσικό για τη συγκέντρωση στοιχείων από clusters από κάθε gmond daemon που περιέχει ολόκληρη την κατάσταση του cluster στο οποίο ανήκει.

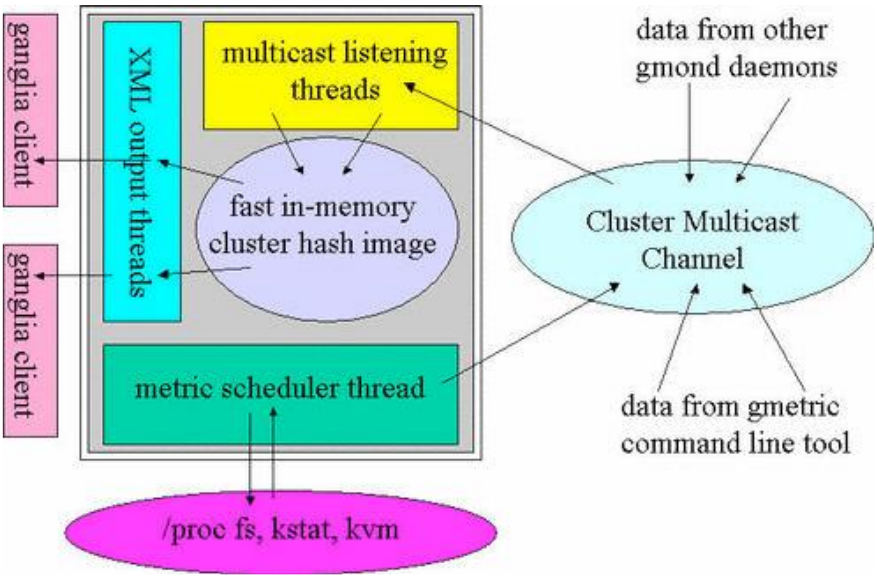
3.7.3 Ganglia PHP Web Front-end

Το Ganglia web front-end παρέχει μια προβολή των πληροφοριών που συλλέγονται μέσω δυναμικών ιστοσελίδων σε πραγματικό χρόνο (Real Time Dynamic Web Pages). Το πιο σημαντικό είναι ότι, εμφανίζει τα δεδομένα του Ganglia με ουσιαστικό και κατανοητό τρόπο για τους διαχειριστές του συστήματος και των χρηστών ηλεκτρονικών υπολογιστών. Παρά το γεγονός ότι το web front-end Ganglia ξεκίνησε ως μια απλή προβολή HTML του δέντρου XML, πλέον έχει εξελιχθεί σε ένα σύστημα που κρατά μια «παλέτα» της ιστορία όλων των δεδομένων που συλλέγονται.

Το Ganglia web front-end μπορεί να ικανοποιήσει τους διαχειριστές του συστήματος και των χρηστών. Για παράδειγμα, μπορεί κανείς να δει τη χρήση της CPU κατά την τελευταία ώρα, ημέρα, εβδομάδα, μήνα ή έτος. Το web front-end δείχνει παρόμοια διαγράμματα για τη χρήση της μνήμης, τη χρήση του δίσκου, τα στατιστικά στοιχεία του δικτύου, τον αριθμό της λειτουργίας των διαδικασιών (processes), καθώς και όλες τις άλλες μετρήσεις του Ganglia.

Το web front-end εξαρτάται από την ύπαρξη του gmetad που του παρέχει δεδομένα από διάφορες πηγές Ganglia. Συγκεκριμένα, το web front-end θα ανοίξει την τοπική θύρα 8651 (από προεπιλογή) και αναμένει να λάβει ένα δέντρο Ganglia XML. Οι ίδιες ιστοσελίδες είναι ιδιαίτερα δυναμικές, οποιαδήποτε αλλαγή στα δεδομένα Ganglia εμφανίζεται αμέσως στην ιστοσελίδα. Αυτή η συμπεριφορά οδηγεί σε ένα site με μεγάλη πληροφόρηση, αλλά απαιτεί το πλήρες δέντρο XML να αναλυθεί σε κάθε πρόσβαση της σελίδας. Ως εκ τούτου, το Ganglia web front-end θα πρέπει να τρέχει σε ένα αρκετά ισχυρό, αφιερωμένο μηχάνημα-υπολογιστή εάν παρουσιάζει μια μεγάλη ποσότητα δεδομένων.

Το Ganglia web front-end είναι γραμμένο σε PHP, και χρησιμοποιεί γραφήματα που δημιουργούνται από το gmetad για να εμφανίσει την ιστορία των πληροφοριών. Έχει δοκιμαστεί σε πολλές εκδόσεις του Unix (κυρίως Linux) με τον server Apache και την ενότητα της PHP 4.1.



Σχήμα 3.5 Γραφική αναπαράσταση του τρόπου επικοινωνίας του Ganglia με το Cluster

3.7.4 Σύνοψη και βασικοί όροι

- Επεκτάσιμο σύστημα παρακολούθησης του cluster (SCMSWeb/SCE Roll).
    - Βασισμένο σε IP multi-cast

- Gmon daemon σε κάθε κόμβο (node) του συστήματος.
    - Multicast κατάσταση συστήματος
    - «Ακούει» άλλα daemons
    - Όλα τα δεδομένα αναπαρίστανται σε XML μορφή.

- Ganglia γραμμή εντολών (comand line)
    - Python code για την ανάλυση της XML σε Αγγλικά

- Gmetric
    - Επεκτάσεις του Ganglia
    - Γραμμή εντολών σε multicast ξεχωριστές μετρήσεις

Αναφορές και Βιβλιογραφία

Πρωτεύουσες αναφορές :

- [1] "About Rocks Cluster". Rocks Cluster Distribution. Retrieved 2011-10-10. http://www.rocksclusters.org/rocks-documentation/4.1/getting-started.html
- [2] "Award Abstract #0438741 SCI: Delivering Cyberinfrastructure: From Vision to Reality". National Science Foundation. July 1, 2005. Retrieved 2012-08-05. https://en.wikipedia.org/wiki/Rocks_Cluster_Distribution
- [3] "Award Abstract #0721623 SDCI: NMI: Improvement: The Rocks Cluster Toolkit and Extensions to Build User-Defined https://en.wikipedia.org/wiki/ Rocks_Cluster_ Distribution
- [4] "SDSC Enhances Rocks Cluster Management Toolkit". Grid Today. February 16, 2004. Archived from the original on 2007-09-27. Retrieved 2012-08-05. https://en.wikipedia.org/wiki/Rocks_Cluster_Distribution
- [5] "Rocks Cluster Register". 2012-11-15. http://www.rocksclusters.org/ rocks-register/
- [6] Gropp, Lusk & Skjellum 1996, p. 3 2012-08-05 https://en.wikipedia.org/wiki/Message _Passing_Interface#I.2FO
- [7] High-performance and scalable MPI over InfiniBand with reduced memory usage.2013-06-16 http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber= Fieeexplore.ieee.org %2Fxpls%2Fabs_all.jsp%3Farnumber%3D4090187
- [8] Table of Contents — September 1994, 8 (3-4). Hpc.sagepub.com. Retrieved on 2014-03-24. https://en.wikipedia.org/wiki/Message_Passing_Interface
- [9] MPI Documents. Mpi-forum.org. Retrieved on 2014-03-24. https://www.mpi-forum.org/.
- [10] Gropp, Lusk & Skjellum 1999b, pp. 4–5 2014-04-08 https://en.wikipedia.org/wiki /Message_Passing_Interface#I.2FO.
- [11] MPI: A Message-Passing Interface Standard Version 3.0, Message Passing Interface Forum, 2012-9-21. https://en.wikipedia.org/wiki/Message_Passing_Interface#I.2FO.
- [12] Russel Sandberg, David Goldberg, Steve Kleiman, Dan Walsh, Bob Lyon (1985). "Design and Implementation of the Sun Network Filesystem". USENIX. 2014-06-11 https://en.wikipedia.org/wiki/Network_File_System
- [13] RFC 3411 — An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks. http://www.rfc-base.org/rfc-3411.html.

❖ [14] Ganglia Monitoring System 2014-5-25. http://ganglia.info/

Δευτερεύουσες αναφορές :

❖ ganglia.info 2014-6-21. http://ganglia.info/

❖ ganglia.sourceforge.net 2014-6-22. http://ganglia.sourceforge.net/

❖ Snir, Marc; Otto, Steve; Huss-Lederman, Steven; Walker, David; Dongarra, Jack (1995) MPI: The Complete Reference. MIT Press Cambridge, MA, USA. ISBN 0-262-69215-5. https://en.wikipedia.org/wiki/Message_Passing_Interface

❖ M Snir, SW Otto, S Huss-Lederman, DW Walker, J (1998) MPI—The Complete Reference: Volume 1, The MPI Core. MIT Press, Cambridge, MA. ISBN 0-262-69215-5. https://mitpress.mit.edu/books/mpi-complete-reference-0

❖ Gropp, William; Steven Huss-Lederman, Andrew Lumsdaine, Ewing Lusk, Bill Nitzberg, William Saphir, and Marc Snir (1998) MPI—The Complete Reference: Volume 2, The MPI-2 Extensions. MIT Press, Cambridge, MA ISBN 978-0-262-57123-4. https://en.wikipedia.org/wiki/Message_Passing_Interface

❖ NFS Illustrated (2000) by Brent Callaghan - ISBN 0-201-32570-5. https://en.wikipedia.org/wiki/Network_File_System
Douglas Mauro, Kevin Schmidt (2005). Essential SNMP, Second Edition. O'Reilly Media. p. 462. ISBN 0596008406. https://en.wikipedia.org/wiki/Network_File_System
Pacheco, Peter S. (1997) Parallel Programming with MPI.[1] 500 pp. Morgan Kaufmann ISBN 1-55860-339-5. https://en.wikipedia.org/wiki/Message_Passing_Interface

❖ Gropp, William; Lusk, Ewing; Skjellum, Anthony (1999b). Using MPI-2: Advanced Features of the Message Passing Interface. MIT Press. ISBN 0-262-57133-1. https://mitpress.mit.edu/books/using-mpi

❖ Gropp, William; Lusk, Ewing; Skjellum, Anthony (1999a). Using MPI, 2nd Edition: Portable Parallel Programming with the Message Passing Interface. Cambridge, MA, USA: MIT Press Scientific And Engineering Computation Series. ISBN 978-0-262-57132-6. https://en.wikipedia.org/wiki/Message_Passing_Interface

❖ Gropp, William; Lusk, Ewing; Skjellum, Anthony (1994). Using MPI: portable parallel programming with the message-passing interface. Cambridge, MA, USA: MIT Press Scientific And Engineering Computation Series. ISBN 0-262-57104-8. https://mitpress.mit.edu/books/series/scientific-and-engineering-computation

❖ Foster, Ian (1995) Designing and Building Parallel Programs (Online) Addison-Wesley ISBN 0-201-57594-9, chapter 8 Message Passing Interface.
https://en.wikipedia.org/wiki/Message_Passing_Interface

ΚΕΦΑΛΑΙΟ 4

Υλοποίηση Cluster σε Εικονικό Περιβάλλον

4.1 Εισαγωγή

Η Υλοποίηση ενός Cluster πρωτίστως εξαρτάται από τον σκοπό για τον οποίο πρόκειται να χρησιμοποιηθεί, δηλαδή τις ανάγκες μας. Κατά δεύτερο τους διαθέσιμους πόρους που έχουμε διαθέσιμους (οικονομικοί, hardware, software, γνώσεις κτλπ). Η δυσκολία της υλοποίησης και πάλι εξαρτάτε από τους προαναφερθέν παράγοντες.

4.2 Εικονικό Cluster

Για να παρακάμψουμε τις όποιες δυσκολίες που σχετίζονται με το hardware, κυρίως τις ασυμβατότητες, ώστε να μπορέσουμε να δούμε πώς δημιουργείται ένα cluster και πως λειτουργεί, η υλοποίηση γίνεται σε εικονικό περιβάλλον (virtual). Αυτό επιτυγχάνεται με την βοήθεια ενός προγράμματος της Oracle με την ονομασία Virtual Box.

4.3 Oracle VM Virtual Box

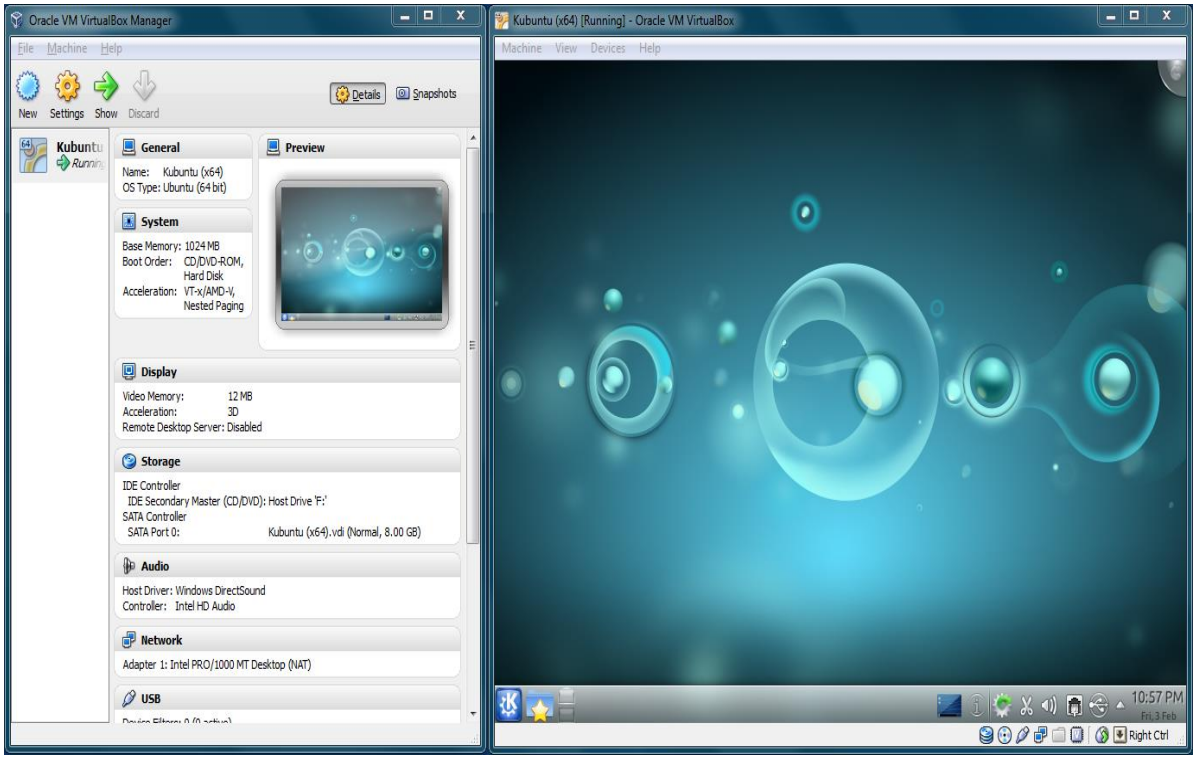
Το Oracle VM VirtualBox είναι ένα πακέτο λογισμικού εικονικοποίησης (virtualization) για συστήματα x86 και AMD64/ Intel64. Το πακέτο VirtualBox εγκαθίσταται σε ένα υπάρχον Λειτουργικό Σύστημα υποδοχής ως εφαρμογή. Αυτή η εφαρμογή υποδοχής επιτρέπει επιπλέον guest λειτουργικά συστήματα, το καθένα είναι γνωστό ως φιλοξενούμενο λειτουργικό. Τέλος μπορεί να φορτώσει και να τρέξει, το καθένα με το δικό του εικονικό περιβάλλον. Υποστηριζόμενα λειτουργικά συστήματα υποδοχής περιλαμβάνουν Linux, Mac OS X, Windows XP, Windows Vista, Windows 7, Windows 8/8.1, Solaris και OpenSolaris.

Οι χρήστες του VirtualBox μπορούν να φορτώσουν πολλαπλά λειτουργικά συστήματα επισκεπτών (guest) σε ένα ενιαίο λειτουργικό σύστημα υποδοχής (host OS). Κάθε επισκέπτης (guest) μπορεί να ξεκινήσει, παγώσει και σταματήσει ανεξάρτητα μέσα στην ίδια εικονική μηχανή (VM). Ο χρήστης μπορεί να ρυθμίσει ανεξάρτητα κάθε VM και να τρέξει κατά επιλογή του λογισμικού(Software) με βάση το virtualization ή υλικό (Hardware) με virtualization, αν υποστηρίζεται αυτό. Στο Λειτουργικό Σύστημα host και το Λειτουργικό Σύστημα guest οι εφαρμογές μπορούν να επικοινωνούν μεταξύ τους μέσω μιας σειράς μηχανισμών, συμπεριλαμβανομένου του κοινού clipboard και μιας εικονικής εγκατάσταση του δικτύου.[1]

Όπως προαναφέρθηκε για την λειτουργία του Virtual Box, είναι απαραίτητο να υπάρχει και υποστήριξη από το Hardware όπου εγκαθίσταται. Γι' αυτό η Intel και η AMD έχουν αναπτύξει η κάθε μία τη δικιά της τεχνολογία για virtualization. Η Intel το VT-x [2]και η AMD το AMD-V [3], όπου ενσωματώνονται στις Κεντρικές Μονάδες Επεξεργασίας (CPU).

4.3.1 Περιορισμοί

- Το Virtual Box δεν υποστηρίζει USB3.
- Το Virtual Box έχει χαμηλή ταχύτητα μεταφοράς από USB συσκευές.
- Ακόμα κι αν το VirtualBox είναι ένα open source προϊόν μερικά από τα χαρακτηριστικά του, παρέχονται μόνο στο πλαίσιο μιας εμπορικής άδειας.
- Περιορισμοί από το hardware του εκάστοτε συστήματος host, σε περίπτωση πολλαπλών virtual guest. (Αδυναμία φόρτωσης ή εκτέλεσης εφαρμογών κλπ.).



Εικόνα 4.1 Screenshot από VirtualBox v4.1.8 «τρέχει» μια έκδοση του Kubuntu 11.04 σε Windows 7 από LiveCD. Στα αριστερά βρίσκεται το VirtualBox virtual machine manager και στα δεξιά η επιφάνεια εργασίας(Desktop) του Kubuntu.

4.4 Εργαλεία για την υλοποίηση

Τα εργαλεία για την υλοποίηση του εικονικού Cluster χωρίζονται σε Software και Hardware όπως θα γινότανε και σε ένα πραγματικό Cluster (εκτός του Oracle Virtual Box). Για να υλοποιηθεί το Cluster χρησιμοποιήθηκαν:

Software:

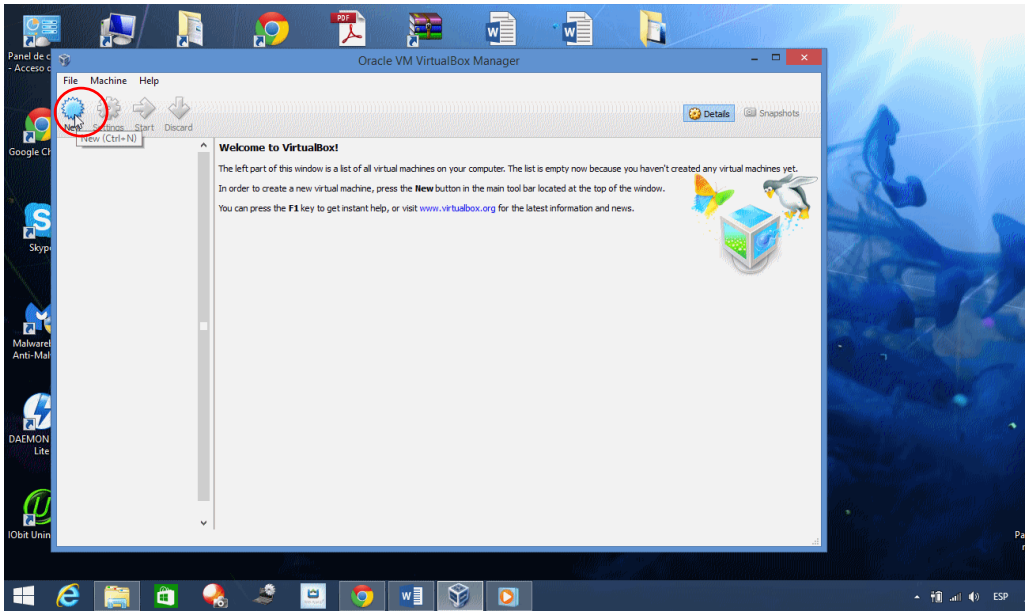
- Oracle VM Virtual Box 4.3.20 Windows hosts x86/amd64
- Λειτουργικό Σύστημα (πελάτη/guest): Rocks Cluster 5.4.3(Viper)
- Λειτουργικό Σύστημα (υποδοχής/host): Windows 8.1 64Bit (ESP)

Hardware:

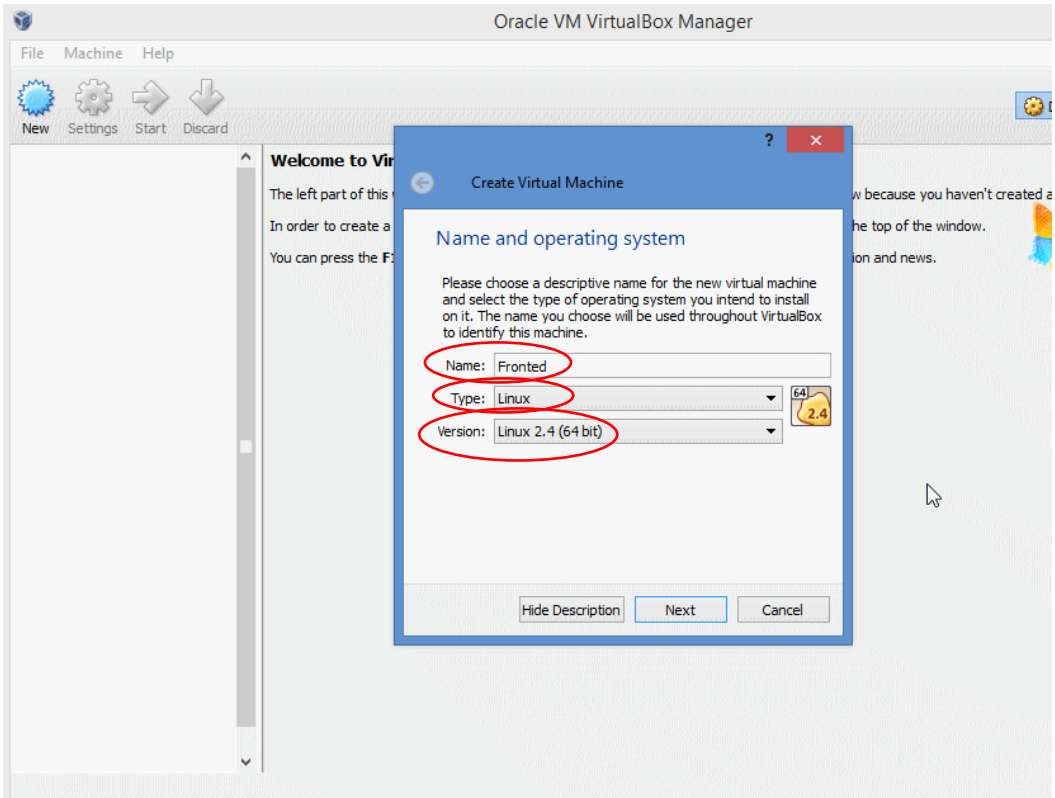
- Laptop: Sony Vaio sve1513c1ew
- CPU: Intel Pentium 2020M 2.4GHz, 2 Cores, VT-x (support)
- Chipset: Intel HM70
- Memory RAM: 8GB DDR3 1600MHz
- GPU: AMD Radeon HD 7650M 1GB
- HDD: Toshiba 500GB, SataII
- Network Adapter Ethernet: Lan Realtek PCIe GBE
- Network Adapter Wifi: Wlan Atheros AR9485WB-EG 802.11n
- DVD RW: Matshita, SataII
- USB 2.0: 3x
- USB 3.0: 1x (δεν υποστηρίζεται από το VirtualBox)

4.5 Υλοποίηση του Cluster βήμα βήμα

Χρησιμοποιώντας τα παραπάνω υλικά , φτιάχνουμε βήμα βήμα το δικό μας Cluster, όπως παρουσιάζεται και στις επόμενες φωτογραφίες-Screenshots. Όπου πάρθηκαν κατά την δημιουργία του Cluster.

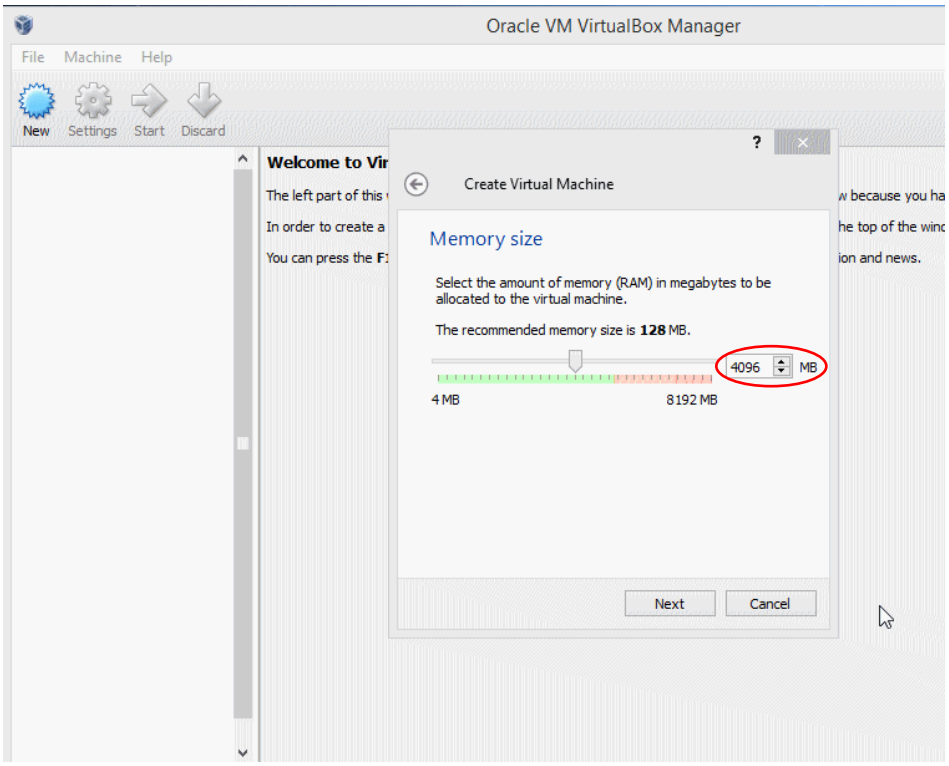


Εικόνα 4.2 Έναρξη Oracle VirtualBox.

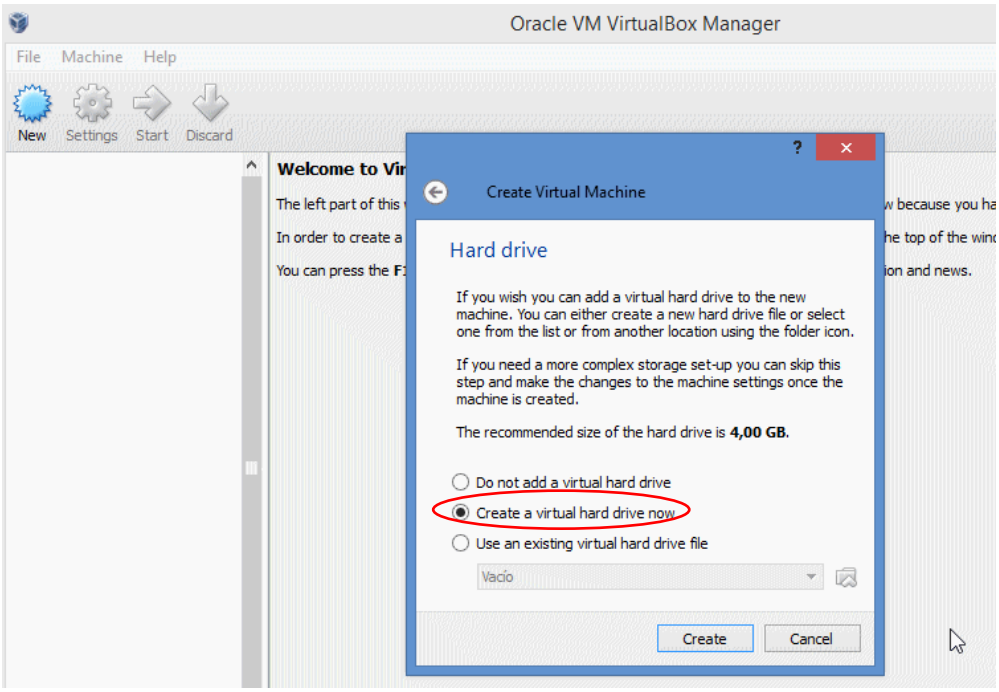


Εικόνα 4.3 Έναρξη δημιουργίας Cluster στο VirtualBox.

Όπως παρουσιάζεται και στα παραπάνω screenshots, ανοίγουμε το VirtualBox, επιλέγουμε ¨New¨ και στην επόμενη καρτέλα (εικόνα 4.3), επιλέγουμε το όνομα (πχ. Fronted), το Λειτουργικό Σύστημα (Linux) και το τύπο-έκδοση του Λειτουργικού Συστήματος που θα χρησιμοποιήσουμε (Linux 2.4 64Bit).

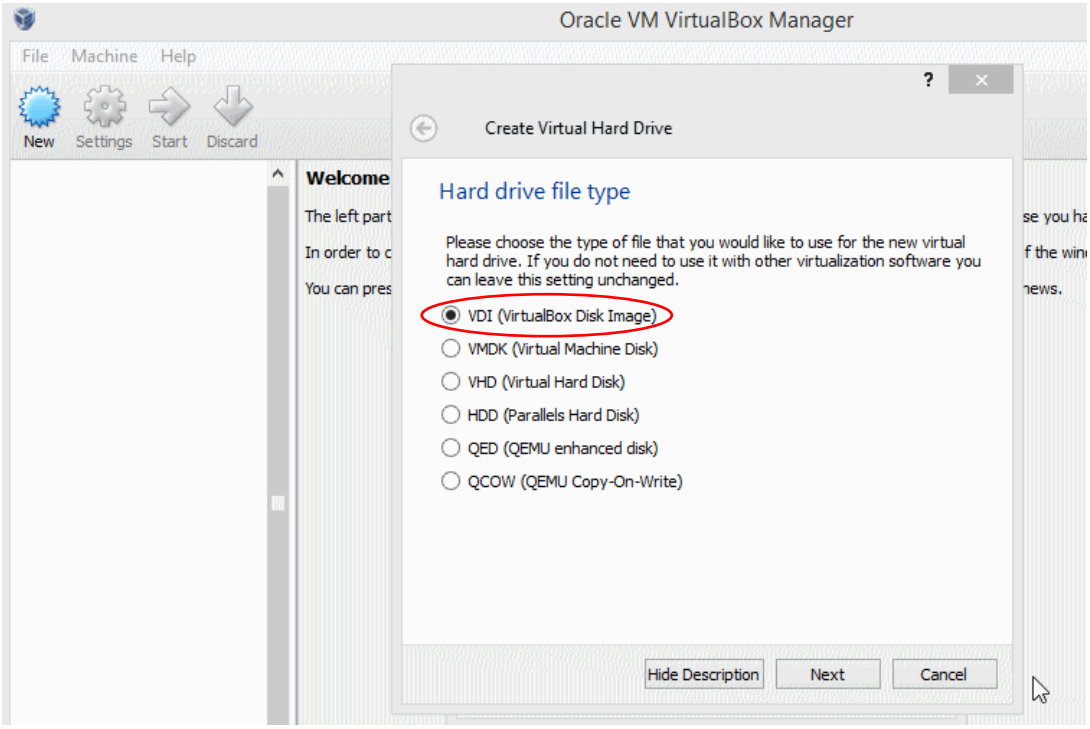


Εικόνα 4.4 Καθορισμός μεγέθους μνήμης

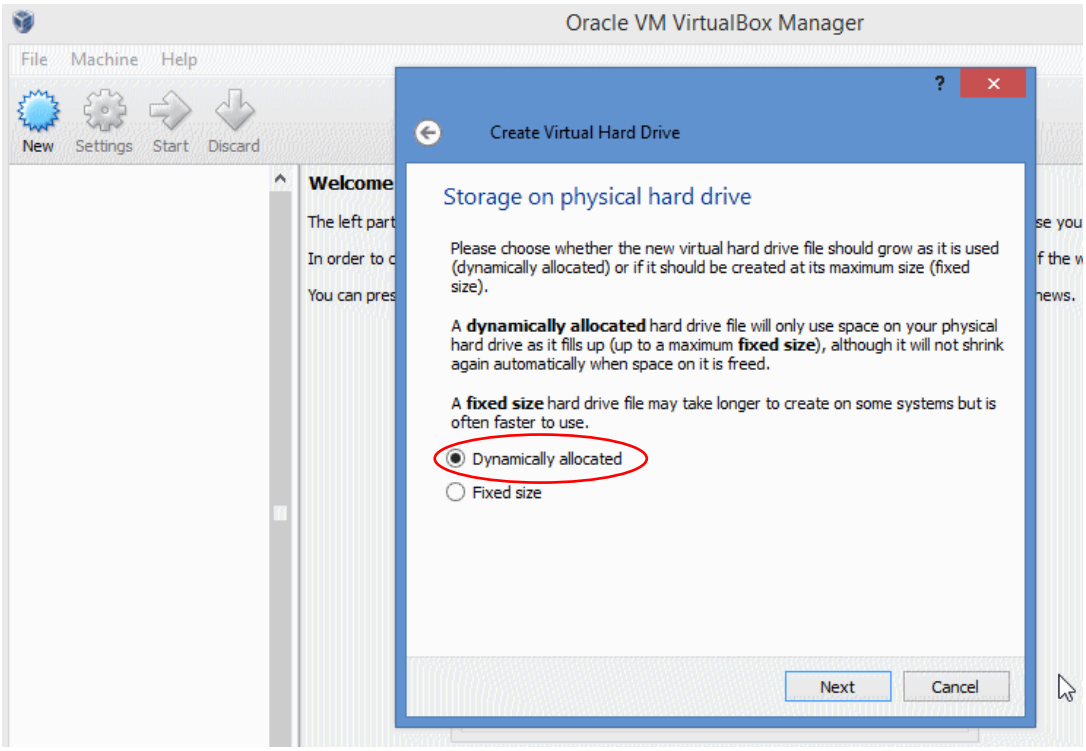


Εικόνα 4.5 Δημιουργία εικονικού σκληρού δίσκου (Virtual Hard Drive).

Στη συνέχεια όπως βλέπουμε και στην Εικόνα 4.4 καθορίζουμε το μέγεθος της μνήμης RAM που θέλουμε να έχει το σύστημα ¨Fronted¨ (πχ. 4GB), ελάχιστη τιμή 512MB ή 1GB για σωστή λειτουργία. Στη συνέχεια δημιουργούμε έναν εικονικό σκληρό δίσκο με ελάχιστη τιμή 16GB.

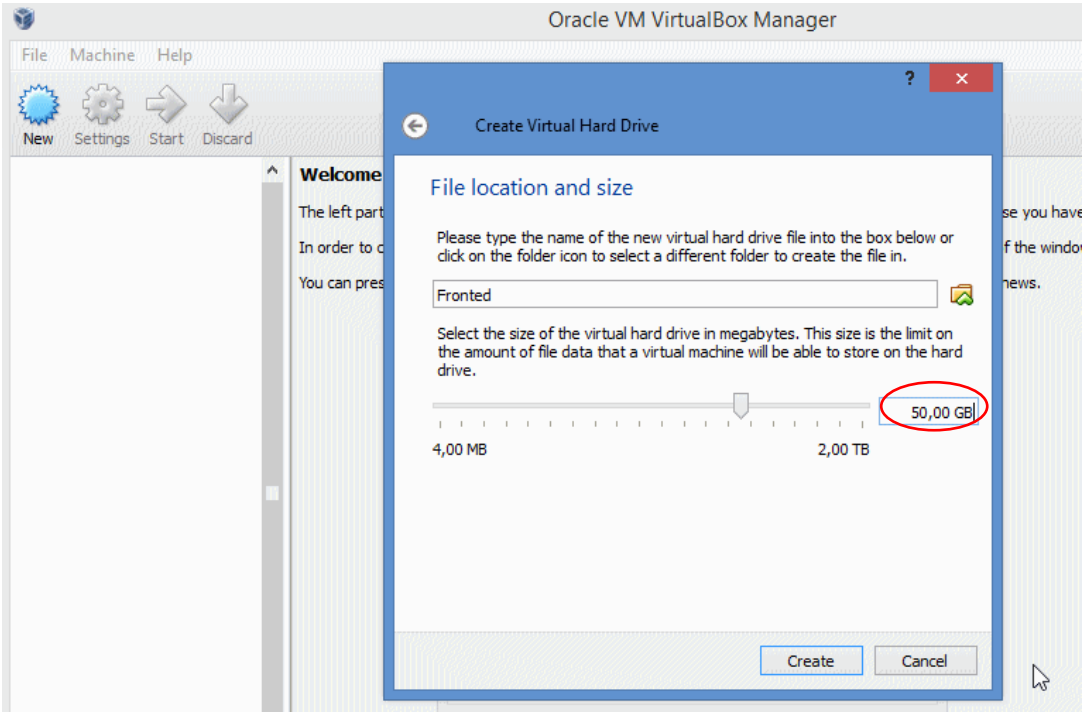


Εικόνα 4.6 Καθορισμός τύπου αρχείου εικονικού σκληρού δίσκου.

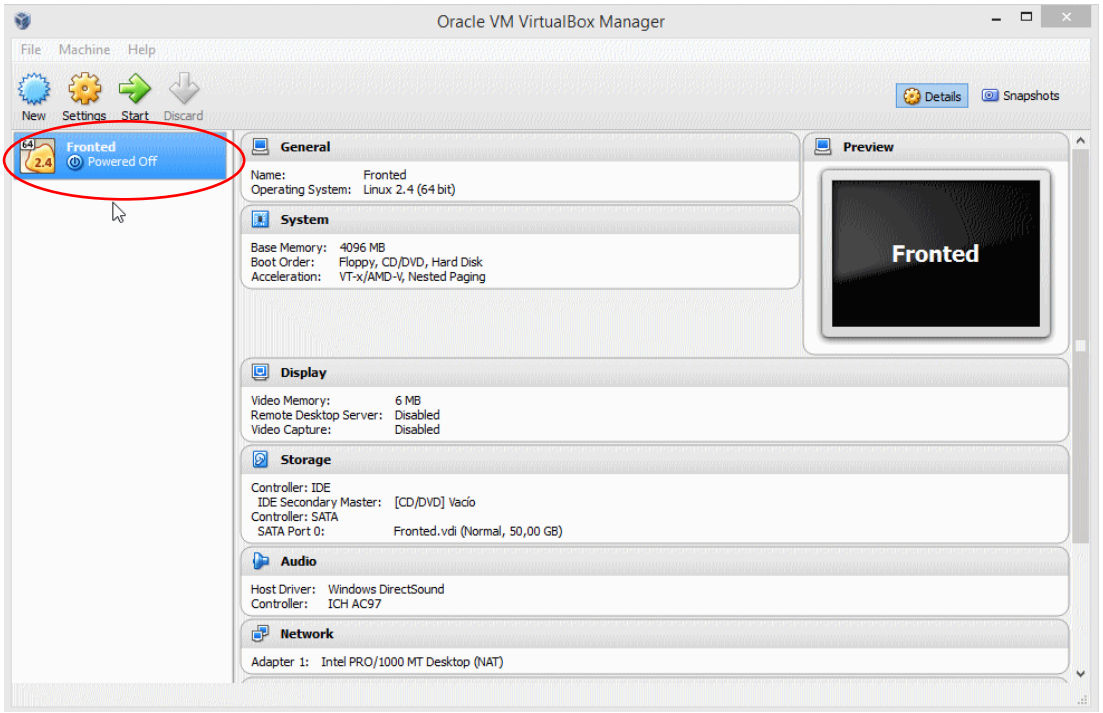


Εικόνα 4.7 Καθορισμός Δυναμικού ή προκαθορισμένου μεγέθους του σκληρού δίσκου.

Στο επόμενο βήμα όπως βλέπουμε από την Εικόνα 4.6 καθορίζουμε τον τύπο του αρχείου όπου θα χρησιμοποιεί ο εικονικός σκληρός δίσκος (Επιλογή ¨VDI VirtualBox Image¨). Έπειτα όπως βλέπουμε στην Εικόνα 4.7 καθορίζουμε τον τρόπο δημιουργίας του εικονικού σκληρού δίσκου επί του φυσικού. (Επιλογή ¨Dynamically¨).

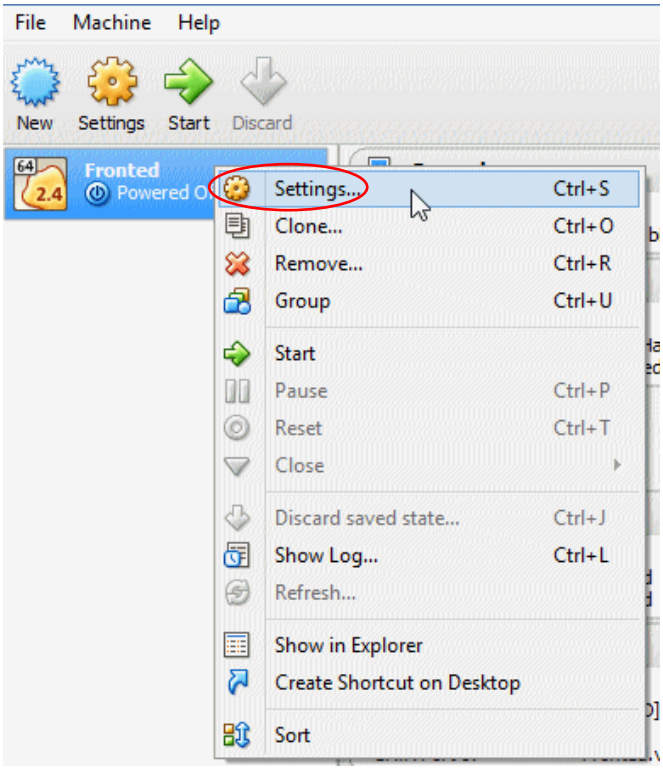


Εικόνα 4.8 Καθορισμός μεγέθους εικονικού σκληρού δίσκου.

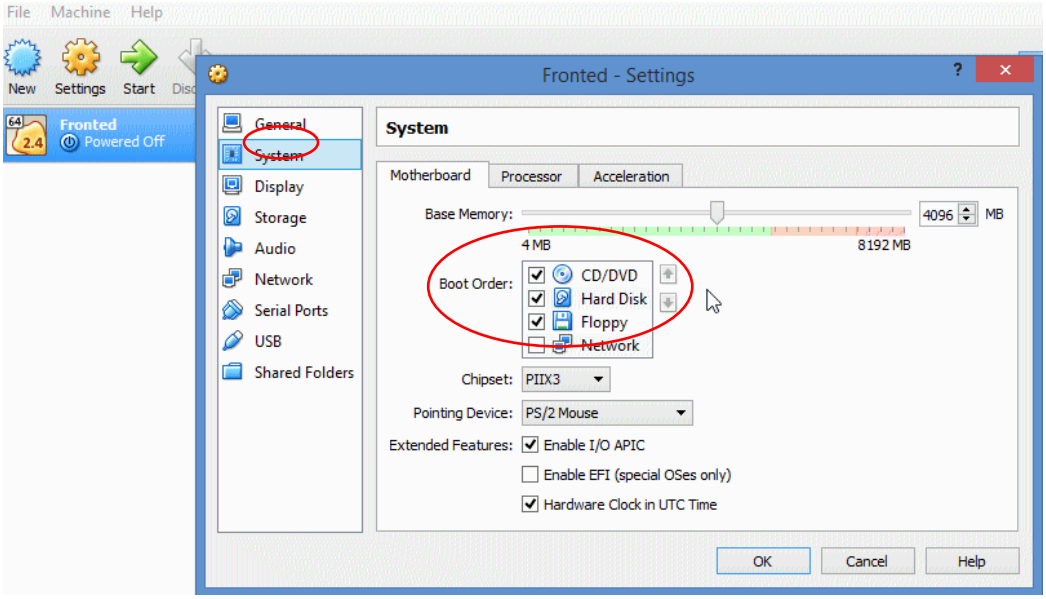


Εικόνα 4.9 Εμφάνιση εικονικού υπολογιστή ¨Fronted¨

Όπως μπορούμε να δούμε στο επόμενο βήμα (Εικόνα 4.8) καθορίζουμε το μέγεθος του εικονικού σκληρού δίσκου έχοντας υπόψη το ελάχιστο όριο (16GB) και τις ανάγκες μας (Επιλογή 50GB). Τέλος βλέπουμε την εμφάνιση του εικονικού υπολογιστή με ονομασία ¨Fronted¨ (Εικόνα 4.9, πάνω αριστερά), καθώς επίσης και τα χαρακτηριστικά του στα δεξιά της εικόνας.(General, System, Display, Storage, κλπ.).

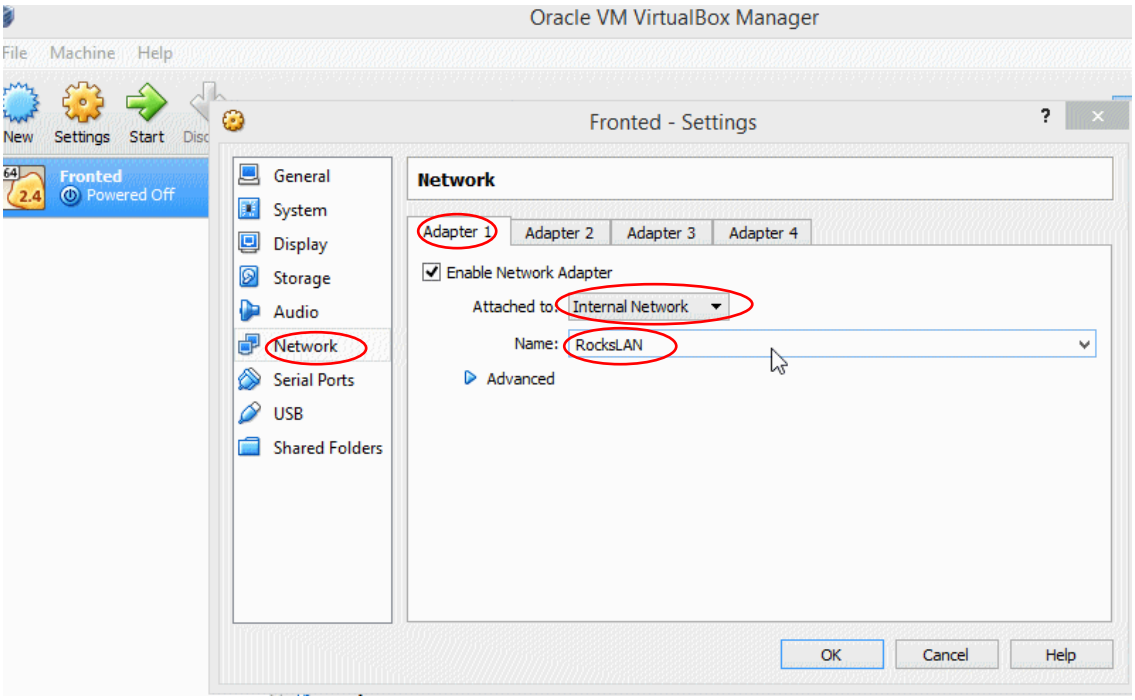


Εικόνα 4.10 Επιλογή ρυθμίσεων ¨Fronted¨

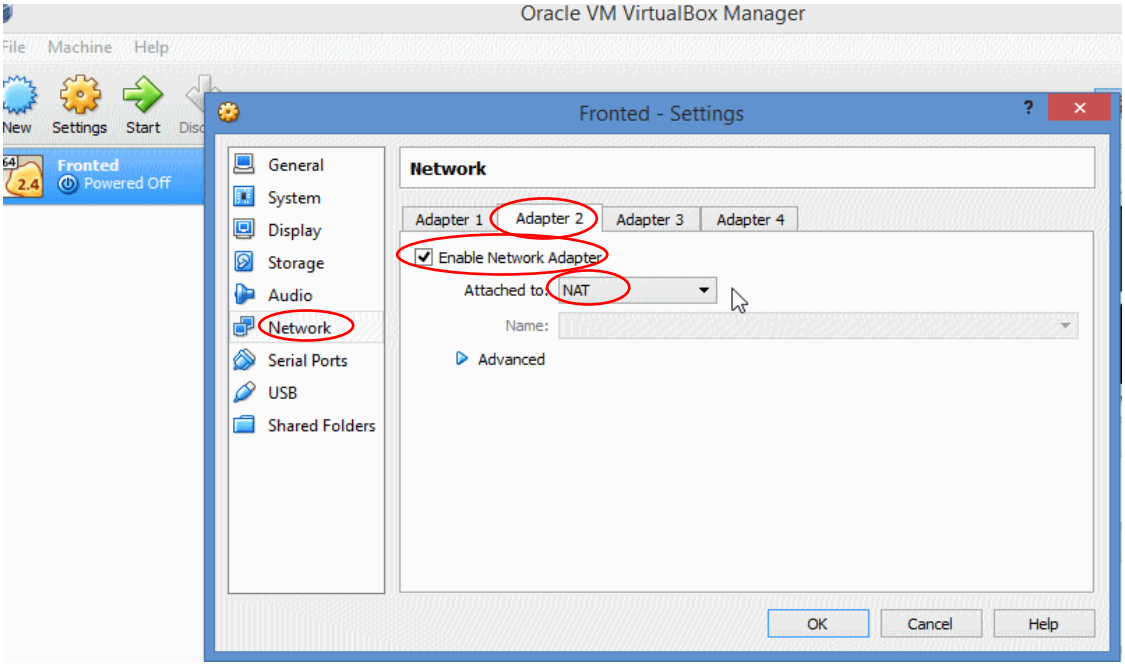


Εικόνα 4.11 Επιλογή προτεραιότητας συσκευών (Boot order)

Εφόσον έχουμε δημιουργήσει το τον εικονικό υπολογιστή ¨Fronted¨, χρειάζεται να προβούμε στις απαραίτητες ρυθμίσεις. Επιλέγουμε δεξί κλίκ ¨Settings" (Εικόνα 4.10) και έπειτα από την καρτέλα ¨System¨ επιλέγουμε την προτεραιότητα των συσκευών κατά την εκκίνηση (Boot order). Ύστερα δίνουμε προτεραιότητα στο CD/DVD, ακολουθεί το HardDisk και τέλος το Floppy.



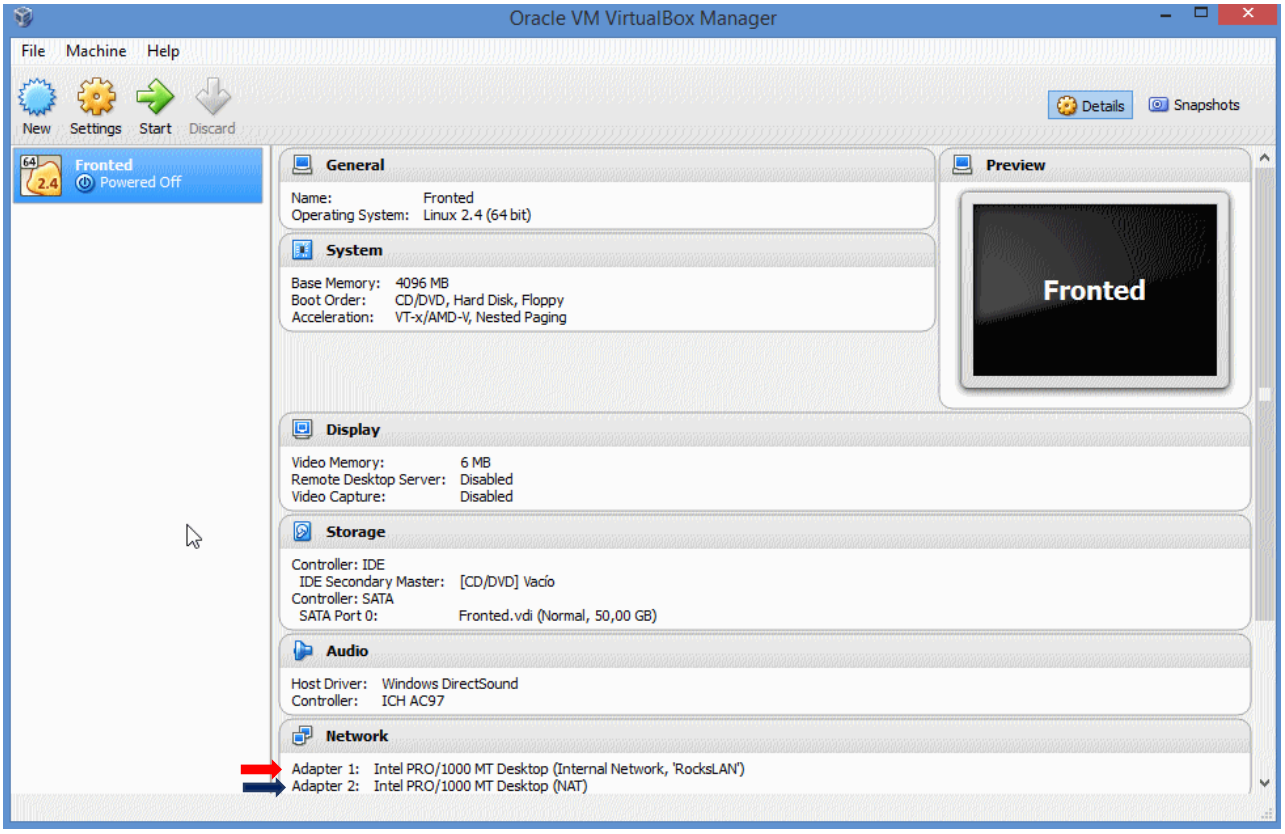
Εικόνα 4.12 Ρυθμίσεις κάρτας δικύου.



Εικόνα 4.13 Ρυθμίσεις δεύτερης κάρτας δικτύου.

Το επόμενο βήμα είναι πολύ καθοριστικό για το σύστημά μας διότι πρέπει να ρυθμίσουμε τις κάρτες δικτύου. Όπως βλέπουμε και στην Εικόνα 4.12 πρώτα ρυθμίζουμε την πρώτη κάρτα δικτύου (adapter 1) και ύστερα επιλέγουμε τον τύπο του δικτύου (Internal Network)και ονομάζουμε το Internal Network (πχ. RocksLAN). Έτσι έχουμε δημιουργήσει μια κάρτα δικτύου αποκλειστικά για το Internal Network. Ύστερα πρέπει να ενεργοποιήσουμε την δεύτερη κάρτα δικτύου(adapter 2) και επιλέγουμε τύπου NAT (network address translation)

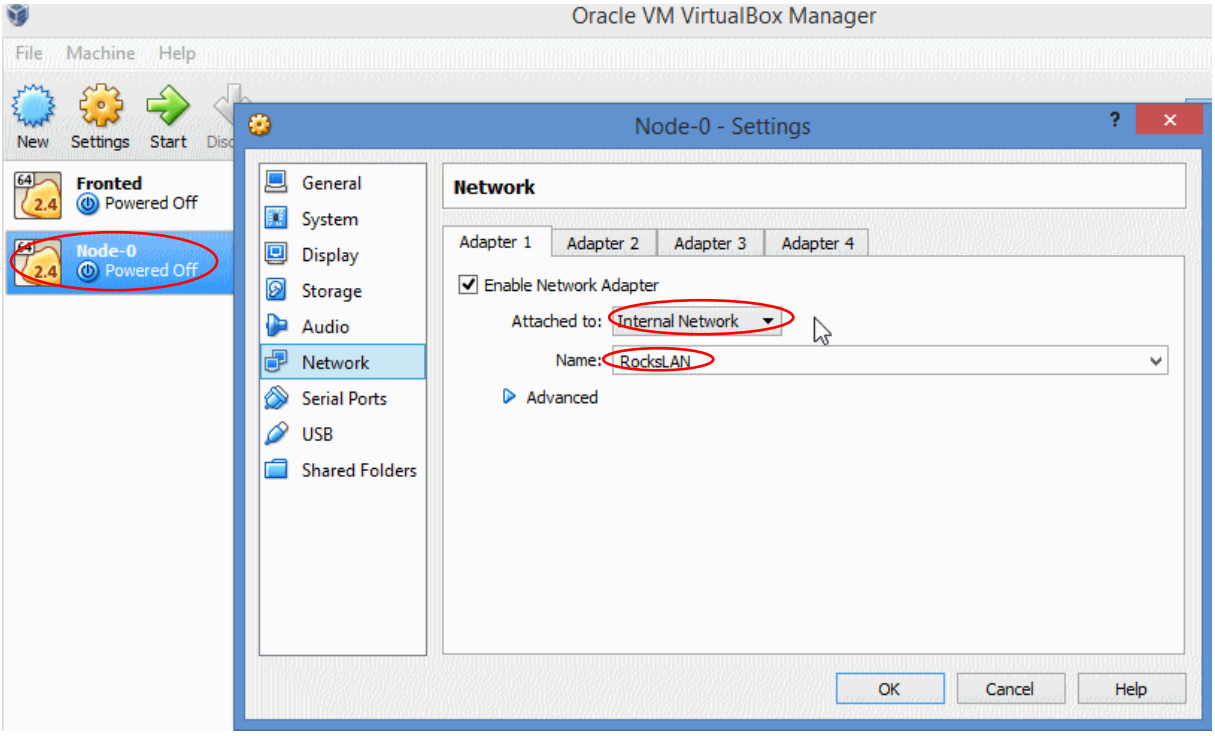
(Εικόνα 4.13). Έτσι έχουμε δημιουργήσει και καθορίσει για κάθε δίκτυο εσωτερικό και εξωτερικό από μία κάρτα δικτύου. Μία για να «μιλάνε» τα nodes-κόμβοι μεταξύ τους (Intranet) και μία για να επικοινωνούν με τον «έξω κόσμο» (Internet).



Εικόνα 4.14 Εμφάνιση των δύο καρτών δικτύων (Με κόκκινο και μπλε βέλος).

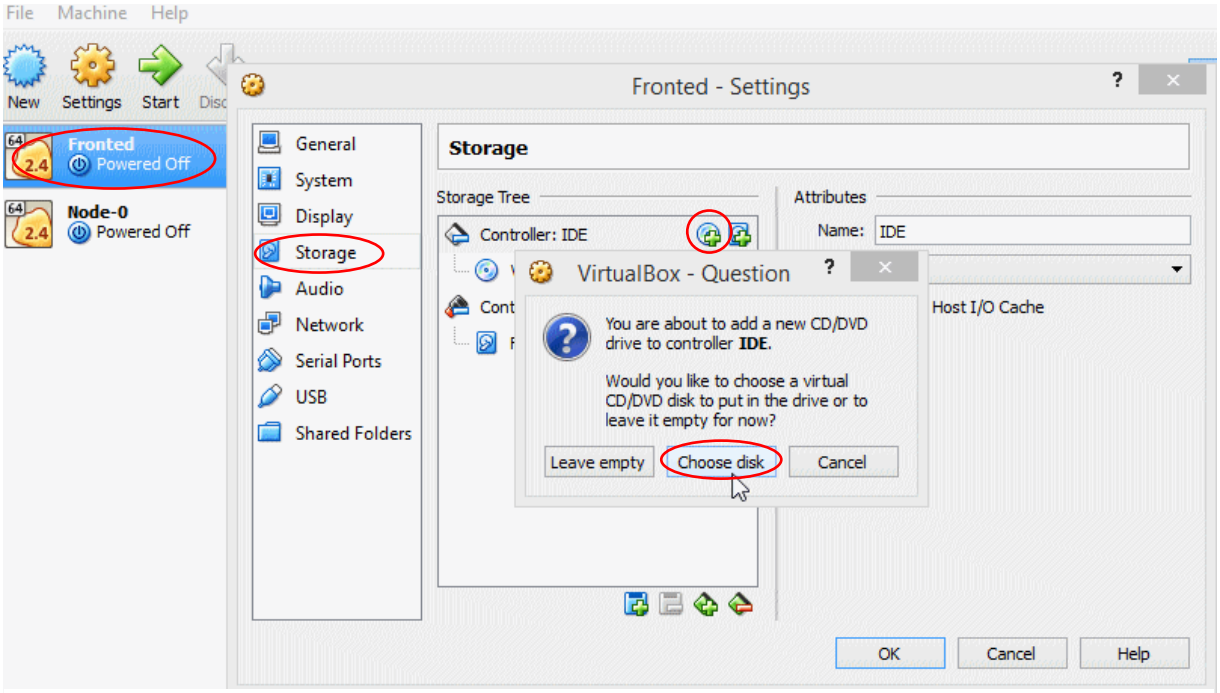
Εφόσον τελειώσαμε και με τη ρύθμιση των καρτών δικτύου του ¨Fronted¨, το επόμενο βήμα είναι να δημιουργήσουμε έναν υπολογιστή node-κόμβο. Η διαδικασία που θα ακολουθηθεί για την κατασκευή του είναι ακριβώς η ίδια με τη δημιουργία του ¨Fronted¨, οι όποιες διαφορές εντοπίζονται στις παραμέτρους της μνήμης RAM , όπου διαθέτει 1GB σε αντίθεση με το ¨Fronted¨ όπου διαθέτει 4GB, επίσης στο μέγεθος του εικονικού σκληρού δίσκου όπου το node διαθέτει 30GB σε αντίθεση με το ¨Fronted¨ όπου διαθέτει 50GB. Αυτή η διαφορά δυναμικής υπάρχει διότι το ¨Fronted¨ είναι συνήθως επιφορτισμένο με περισσότερες λειτουργίες απ' ότι τα υπόλοιπα nodes-κόμβοι. Τέλος μια σημαντική διαφορά μεταξύ ¨Fronted¨ και κόμβου-node, εντοπίζεται στην κάρτα δικτύου, διότι ο κάθε κόμβος-node θα πρέπει να συνδέεται με το ¨Fronted¨ μέσω του Εσωτερικού Δικτύου (Internal Network), ώστε να προσφέρουν τις υπηρεσίες τους στο ¨Fronted¨ καθώς και να επικοινωνούν με τον «έξω κόσμο»

διαμέσου του ¨Fronted¨. Όπως παρουσιάζεται στην Εικόνα 4.15 όπου επιλέξαμε το Internal Network με ονομασία «RocksLAN» για το ¨Node-0¨.



Εικόνα 4.15 Ρυθμίσεις στην κάρτα δικτύου του Node-0.

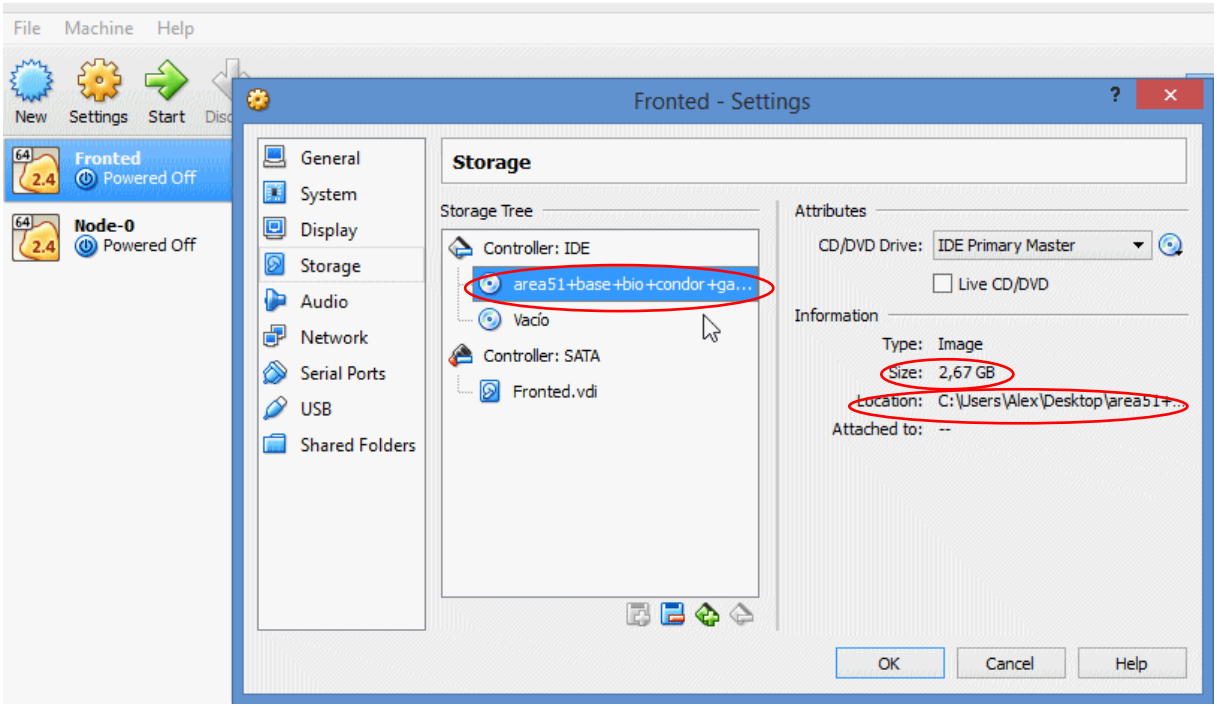
4.5.1 Εγκατάσταση Rocks Cluster στο ¨Fronted¨



Εικόνα 4.16 Επιλογή και φόρτωση εικονικού δίσκου στο CD/DVD του¨Fronted¨.

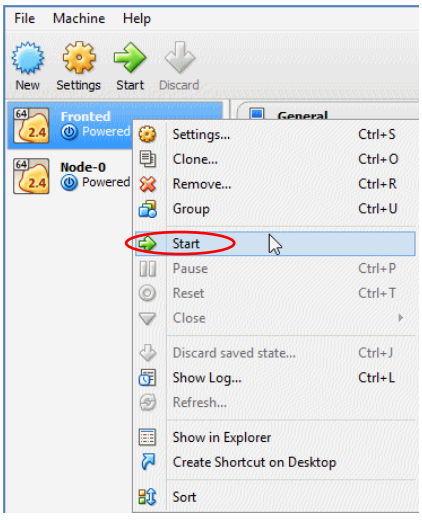
Μετά την επιτυχή δημιουργία του ¨Node-0¨ και των ρυθμίσεών του, επανερχόμαστε στο ¨Fronted¨ και μέσα από τα “Settings”, επιλέγουμε “Storage” και έπειτα από το εικονίδιο με τον

«CD και τον πράσινο σταυρό» επιλέγουμε το “Choose disk”, για να εισάγουμε τον εικονικό δίσκο.

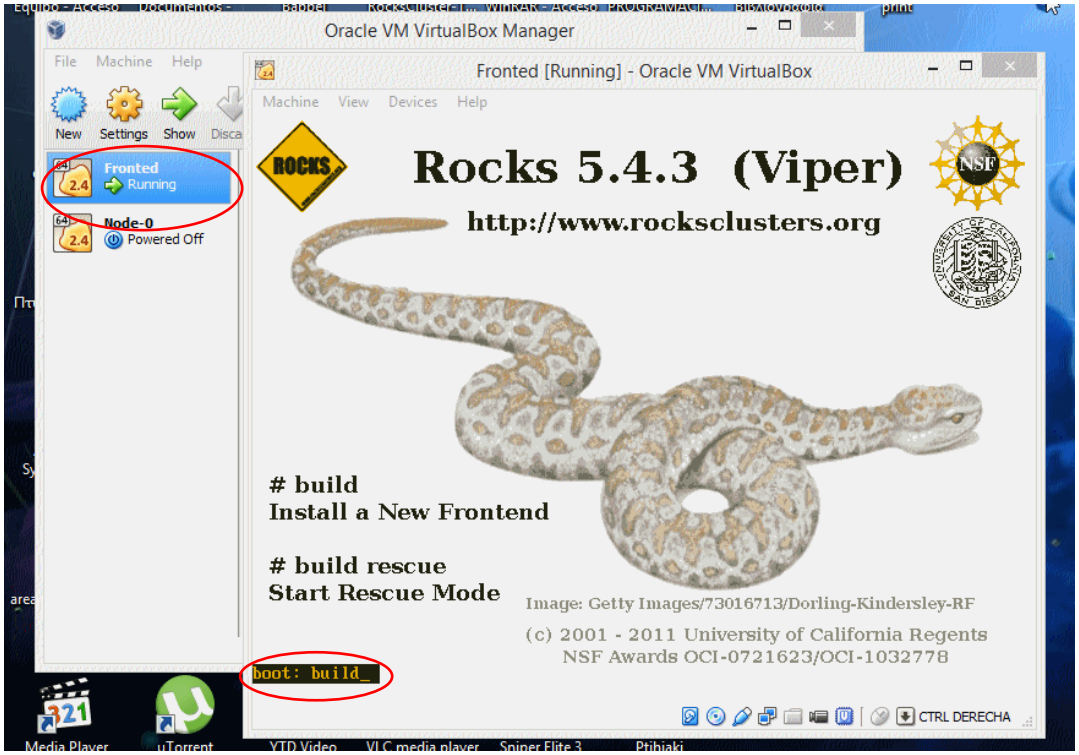


Εικόνα 4.17 Εισαγωγή του εικονικού δίσκου.

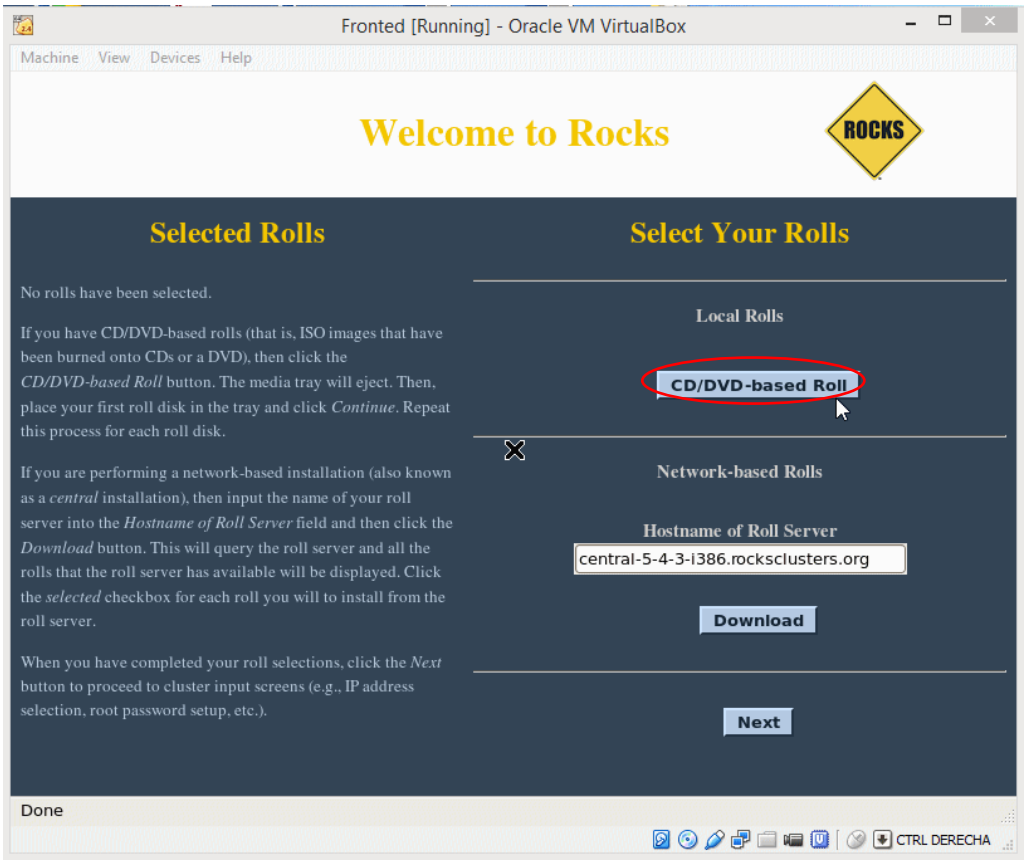
Όπως βλέπουμε και στην Εικόνα 4.17 εισάγουμε τον δίσκο στο CD/DVD του “Fronted”, στην προκειμένη περίπτωση το “RocksCluster v5.4.3 viper”, με μέγεθος όπως βλέπουμε 2.67GB, από την τοποθεσία όπου βρίσκεται αποθηκευμένο, είτε σε φυσικό μέσο CD/DVD είτε τοπικά αποθηκευμένο στον φυσικό σκληρό δίσκο του συστήματός μας σε μορφή “.iso” όπως στην περίπτωσή μας. Ύστερα από τα “Settings¨ του “Fronted” επιλέγουμε “Start” ώστε να εκκινήσουμε τον εικονικό υπολογιστή, όπως φαίνεται και στην εικόνα 4.18.



Εικόνα 4.18 Έναρξη εικονικού υπολογιστή “Fronted”



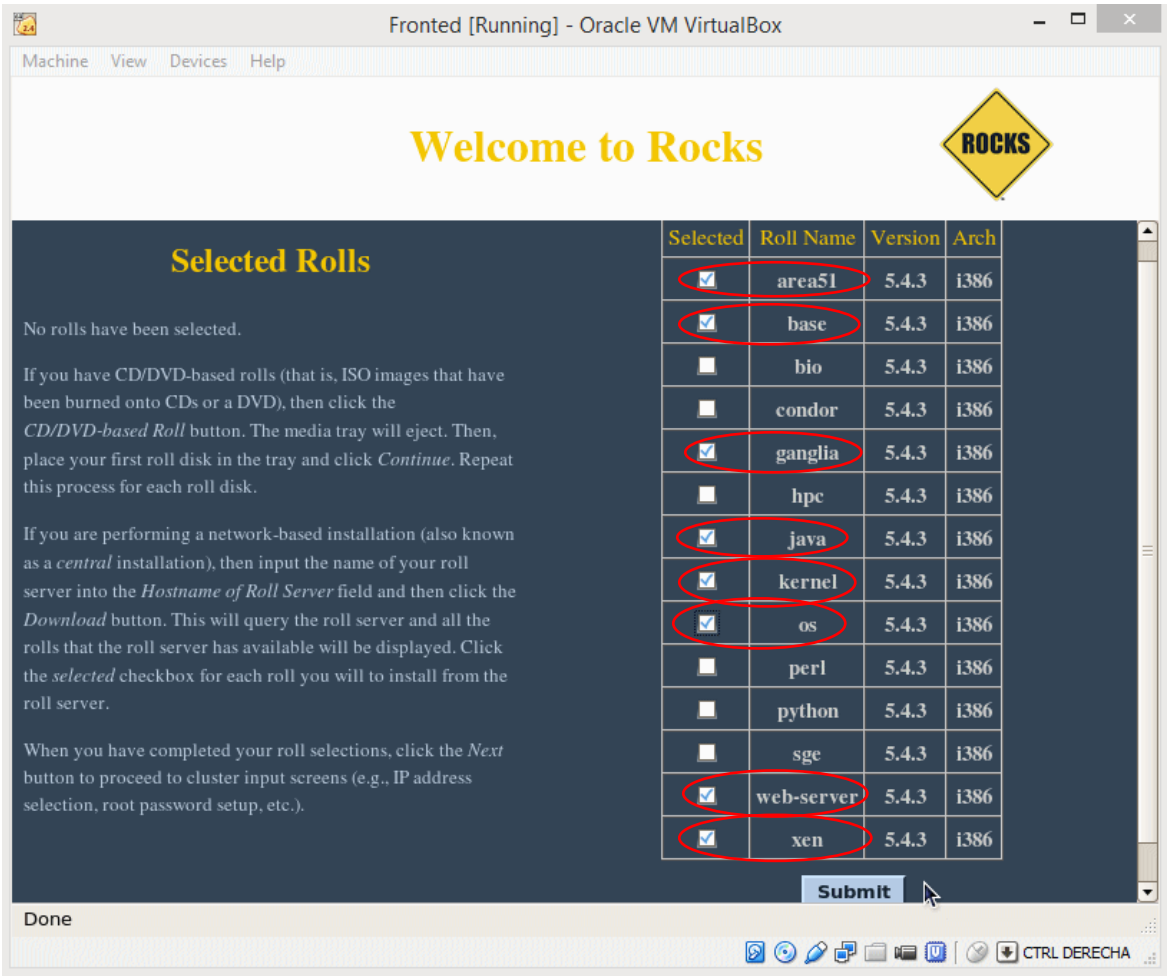
Εικόνα 4.19 Έναρξη εγκατάστασης Rocks Cluster v5.4.3(Viper).



Εικόνα 4.20 Είσοδος στο CD/DVD του based Roll στο Rocks Cluster.

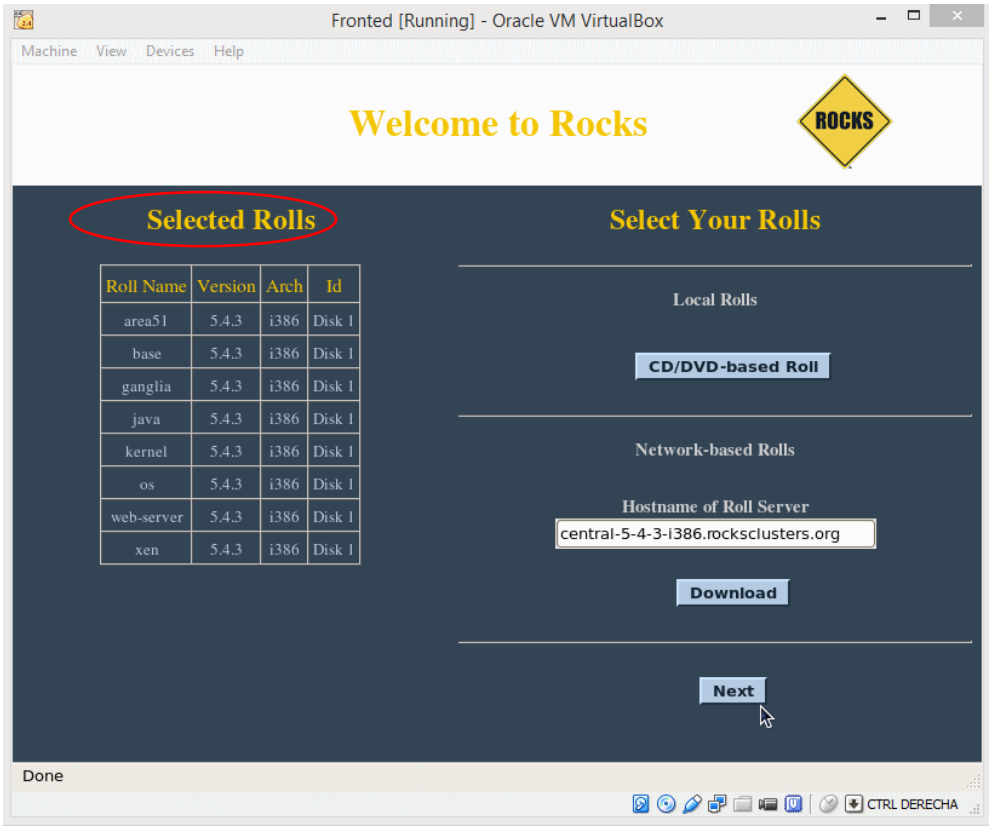
Τα επόμενα βήματα αφορούν στην εγκατάσταση της διανομής RocksCluster v5.4.3 Viper, καθώς και στις απαραίτητες ρυθμίσεις που θα πρέπει να προβούμε. Όπως βλέπουμε και

στην Εικόνα 4.19 με την εντολή ¨build¨ ξεκινάμε την εγκατάσταση, ύστερα πραγματοποιούμε είσοδο στο ¨CD/DVD-based Roll.¨ όπου βρίσκονται όλα τα απαραίτητα «συστατικά» Roll´s, για την δημιουργία του Cluster μας (Εικόνα 4.20).

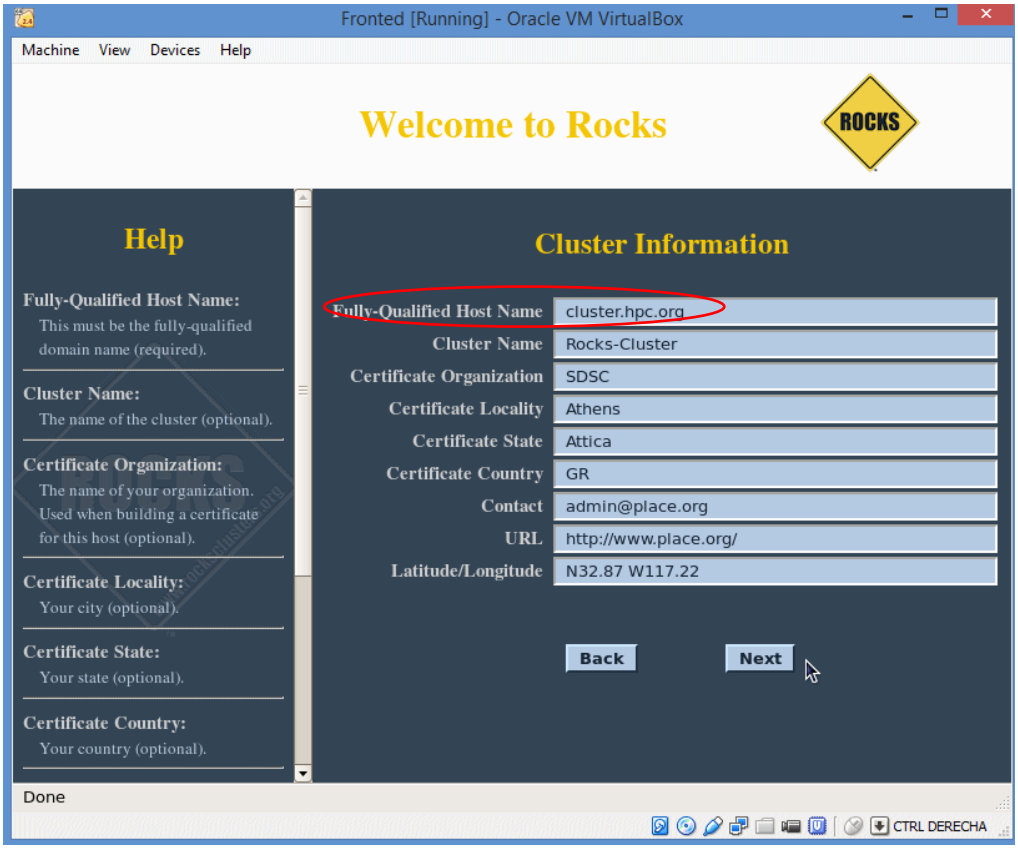


Εικόνα 4.21 Επιλογή των Roll´s

Τα Roll´s CD´s είναι πακέτα επέκτασης των δυνατοτήτων του Λειτουργικού Συστήματος, όπου ο χρήστης μπορεί να τα επιλέξει ανάλογα με τις ανάγκες του και να διαμόρφωση το σύστημά του. Για την δημιουργία του δικού μας συστήματος χρησιμοποιήσαμε όπως φαίνεται και στην Εικόνα 4.21 τα Roll´s CD´s, “area 51”, “base”, “ganglia”, “java”, “kernel”, “os”, “web-server” και“xen”. Ύστερα υπάρχει προεπισκόπηση των επιλογών μας, ώστε να διαπιστώσουμε εάν κάναμε τις επιλογές τις αρεσκείας μας, όπως παρουσιάζεται στην εικόνα 4.22.

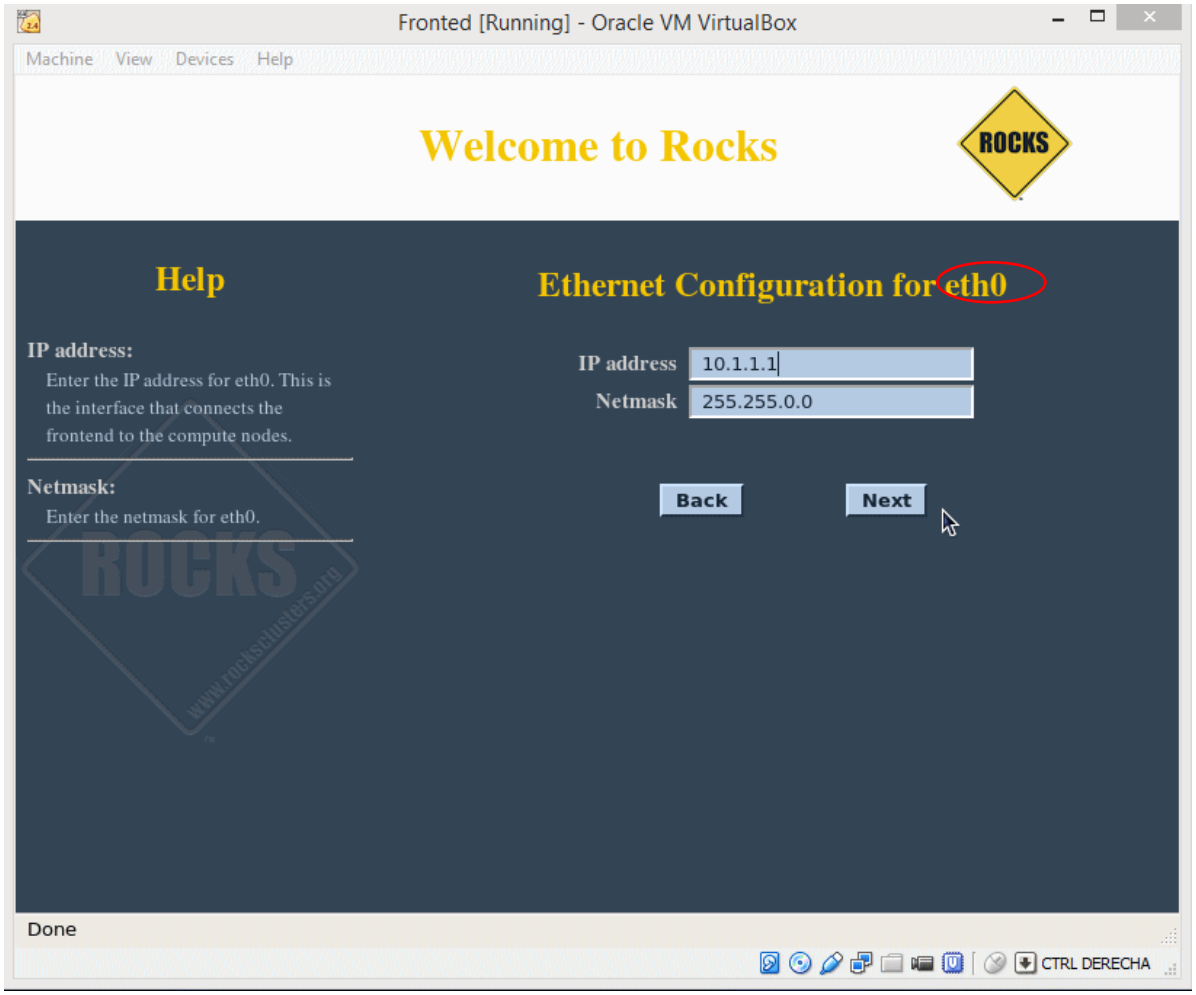


Εικόνα 4.22 Προεπισκόπηση επιλεγμένων Roll΄s CD΄s



Εικόνα 4.23 Πληροφορίες του Cluster.

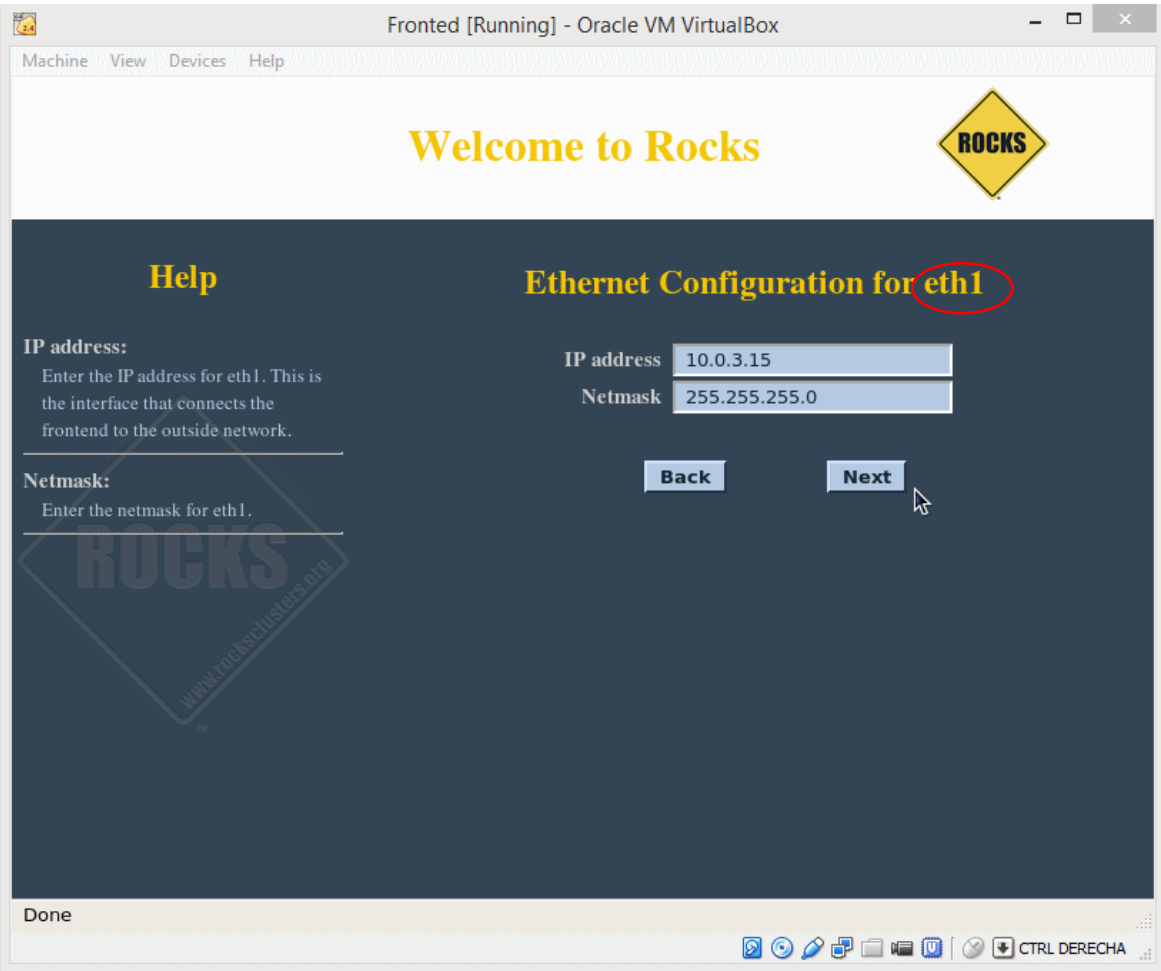
Μετά την προεπισκόπηση των Roll´s CD´s που επιλέξαμε (Εικόνα 4.22), ορίζουμε τις πληροφορίες για το Cluster. Το μοναδικό στοιχείο που είναι απαραίτητο, είναι το ¨host name¨, όπου μπορούμε να το αφήσουμε στην προεπιλογή του, εκτός και αν διαθέτουμε κάποιο δικό μας. Όλα τα υπόλοιπα στοιχεία (όνομα cluster, όνομα πόλης, νομού, χώρας, κλπ.) είναι προαιρετικά (Εικόνα 4.23).



Εικόνα 4.24 Παραμετροποίηση της κάρτας δικτύου Ethernet ¨eth0¨.

Εφόσον ολοκληρωθεί και η φάση της συμπλήρωσης των στοιχείων, ακολουθεί η ρύθμιση των καρτών δικτύου. Πρώτα με την κάρτα δικτύου Ethernet ¨eth0¨, όπου μέσω αυτής οι κόμβοι στο σύστημά μας μπορούν να συνδεθούν με τον υπολογιστή ¨Fronted¨, δηλαδή αναφέρεται στο εσωτερικό δίκτυο όπου έχουμε δημιουργήσει (Internal Network) με την ονομασία όπου του έχουμε δώσει ¨RocksLAN¨. Εδώ ορίζουμε την IP διεύθυνση και την Netmask όπου θα βλέπουν οι κόμβοι-nodes, στο Internal Network (Εικόνα 4.24).Στη συνέχεια ρυθμίζουμε την δεύτερη κάρτα δικτύου όπου ο υπολογιστής ¨Fronted¨ θα μπορεί να

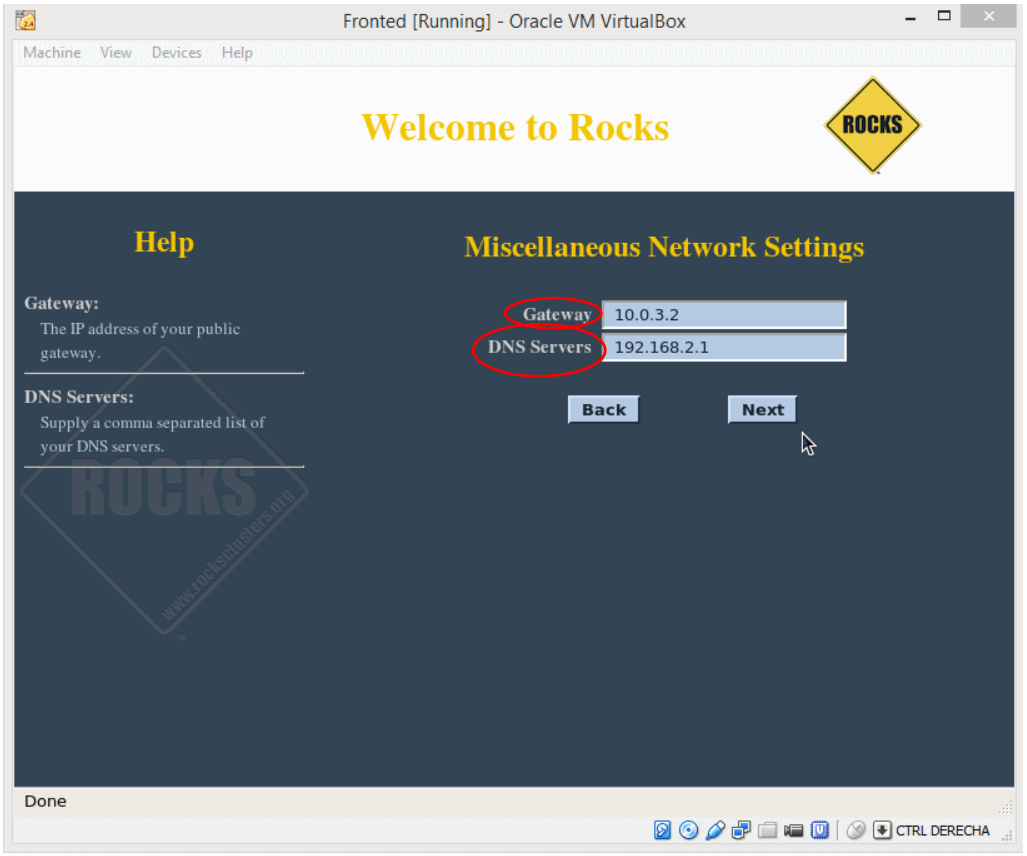
επικοινωνεί με τον «έξω κόσμο» Internet, καθώς και οι κόμβοι-nodes μέσο αυτού ¨eth1¨ (Εικόνα 4.25).



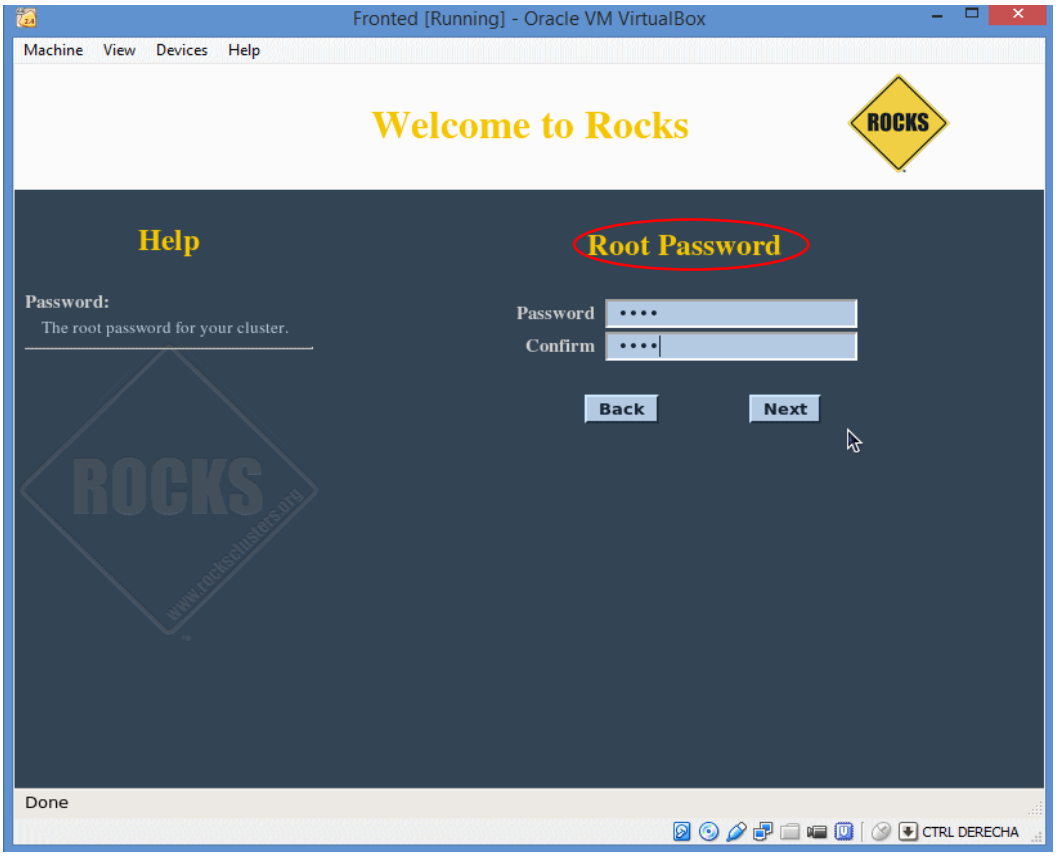
Εικόνα 4.25 Παραμετροποίηση δεύτερης κάρτας δικτύου Ethernet ¨eth1¨.

Ύστερα από την ρύθμιση των δύο καρτών δικτύων, προχωράμε στο να ορίσουμε μια διεύθηνση για το Gateway, όπου θα επιτρέπει το "Fronted" καθώς και όλο το Internal Network να συνδέεται με το Internet. Επίσεις και τον ορισμό ενός DNS Server όπως φένεται στην Εικόνα 4.26 παρακάτω.

Ολοκληρόνωντας και αυτό το βήμα, προχωράμε στο να ορίσουμε έναν κωδικό ασφαλίας, όπου κατά την είσοδό μας στο Cluster θα μας ζητήτε ώστε να μπωρούμε να συνδεθούμε, το ίδιο δηλαδή που θα απετούσε ένα κανονικό Λειτουργικό Σύστημα. Όπως μπωρούμε να δούμε και στην Εικόνα 4.27.

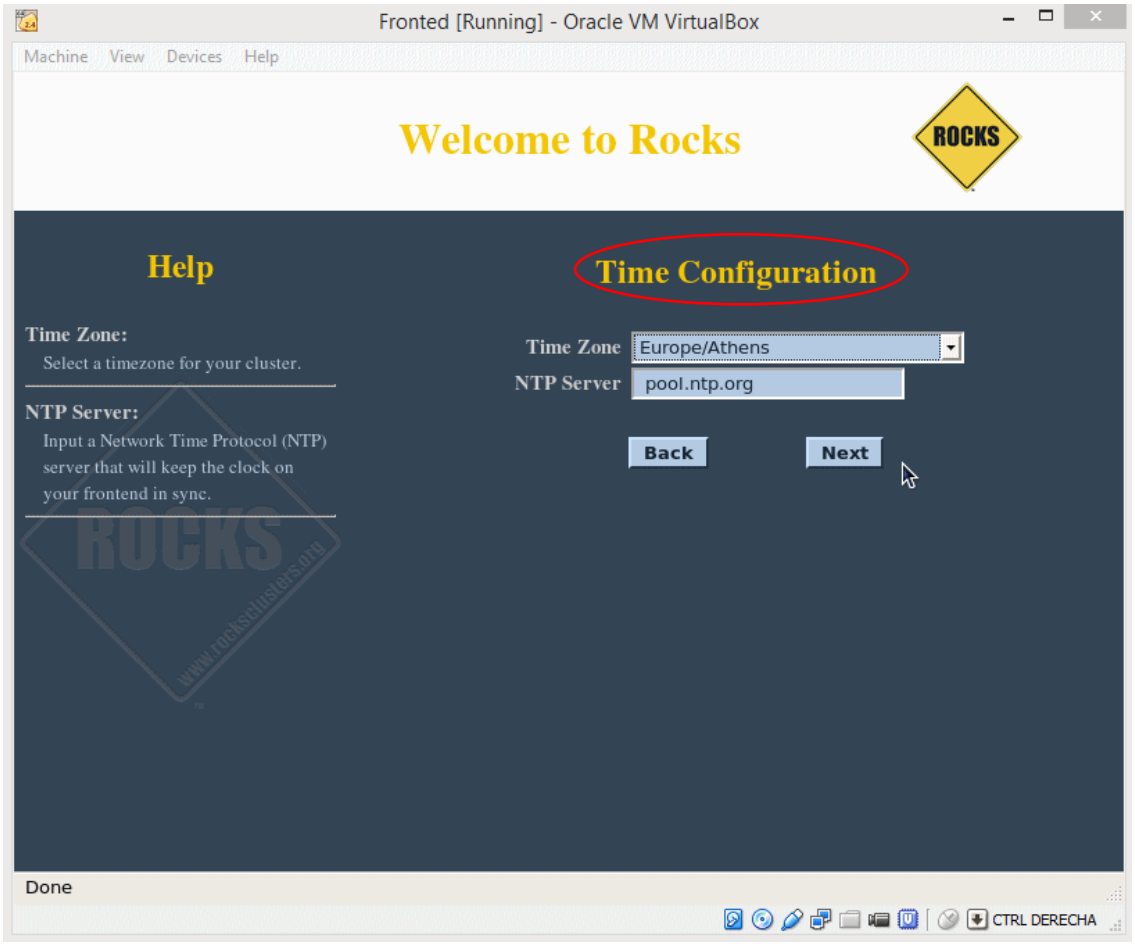


Εικόνα 4.26 Ρύθμιση Gateway και ορισμός DNS Server.



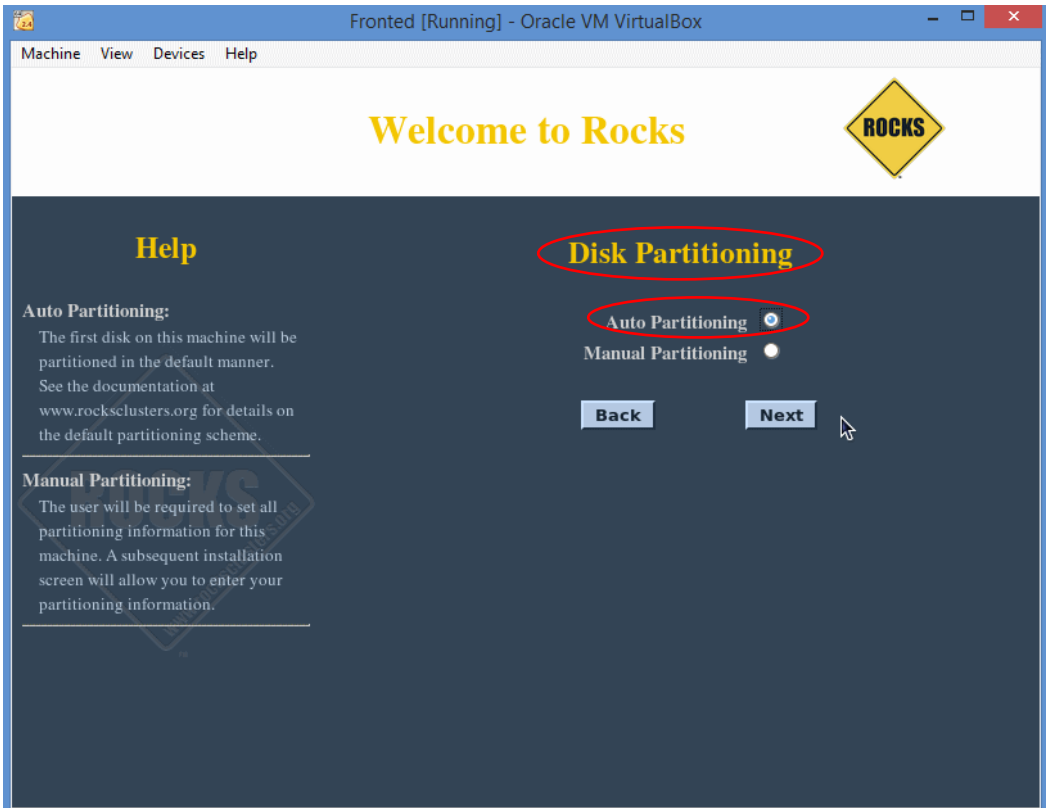
Εικόνα 4.27 Καθορισμός κωδικού ασφαλείας (Password).

Φτάνοντας στο τέλος των απαραίτητων και καθοριστικών ρυθμίσεων για το Cluster μας, στα επόμενα βήματα, ύστερα από τον καθορισμό κωδικού ασφαλείας (Password), μας ζητάτε να προσδιορίσουμε τη ζώνη ώρας στην οποία ανήκουμε, καθώς και NTP Server (Network Time Protocol) όπου θα ρυθμίζει το ρολόι στον υπολογιστή για κρατάει συγχρονισμένο το ¨Fronted¨. (Εικόνα 4.28)



Εικόνα 4.28 Ορισμός ζώνη ώρας και NTP Server (Network Time Protocol).

Τέλος πρέπει να καθοριστεί ο τρόπος εγκατάστασης του Λειτουργικού Συστήματος στον σκληρό δίσκο του συστήματός μας. Εδώ υπάρχουν δύο επιλογές στη διάθεσή μας η μία είναι το ¨Auto Partitioning¨ και η άλλη το ¨Manual Partitioning¨, δηλαδή είτε να χωρίσουμε και να διαμορφώσουμε σε τμήματα (Partitions) το σκληρό δίσκο αυτόματα είτε κατά βούληση. Το ίδιο ζητείτε και όταν θέλουμε να εγκαταστήσουμε ένα οποιοδήποτε Λειτουργικό Σύστημα πχ μια οποιαδήποτε άλλη διανομή Linux ή Windows.(Εικόνα 4.29).Εμείς επιλέξαμε το ¨Auto Partitioning¨.



Εικόνα 4.29 Διαχωρισμός σε τμήματα-Partitions του σκληρού δίσκου.

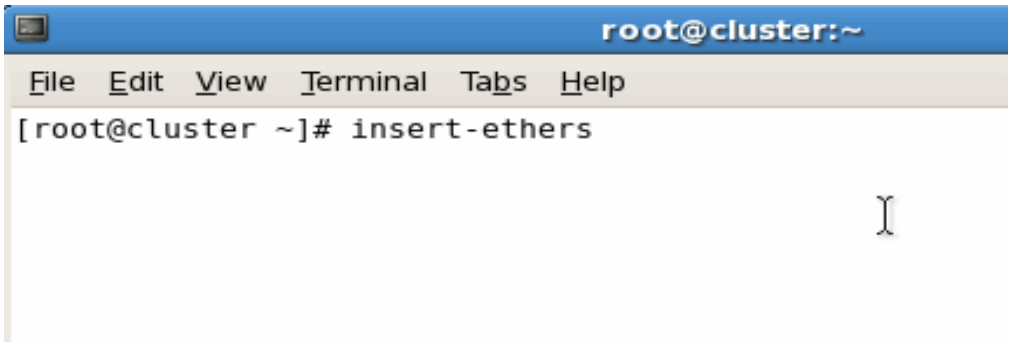
Εικόνα 4.30 Το VirtualBox Machine (αριστερά) μαζί με την επιφάνεια εργασίας Rocks Cluster v.5.4.3 viper (CentOs 5).

Εφόσον ολοκληρωθεί ο καθορισμός των απαραίτητων ρυθμίσεων του Λειτουργικού Συστήματος, η εγκατάσταση θα χρειαστεί κάποιο χρόνο για να ολοκληρωθεί. Όταν η εγκατάσταση ολοκληρωθεί θα μας ζητηθεί το όνομα χρήστη (User Name), καθώς και τον κωδικό ασφαλείας (Password) όπου του έχουμε καθορίσει από τις ρυθμίσεις κατά την

εγκατάσταση. Πραγματοποιώντας τα βήματα αυτά με επιτυχία θα πρέπει να μας εμφανίσει την επιφάνεια εργασίας του Rocks Cluster v5.4.3 viper, όπως στην εικόνα 4.30. Στα αριστερά της εικόνας μπορούμε να διακρίνουμε το Virtual Box machine όπου είναι οι εικονικοί υπολογιστές μας καθώς και τον ¨Fronted¨ να βρίσκετε σε λειτουργία ¨Running¨ και στα δεξιά της εικόνας την επιφάνεια εργασίας του Rocks Cluster v5.4.3 viper.

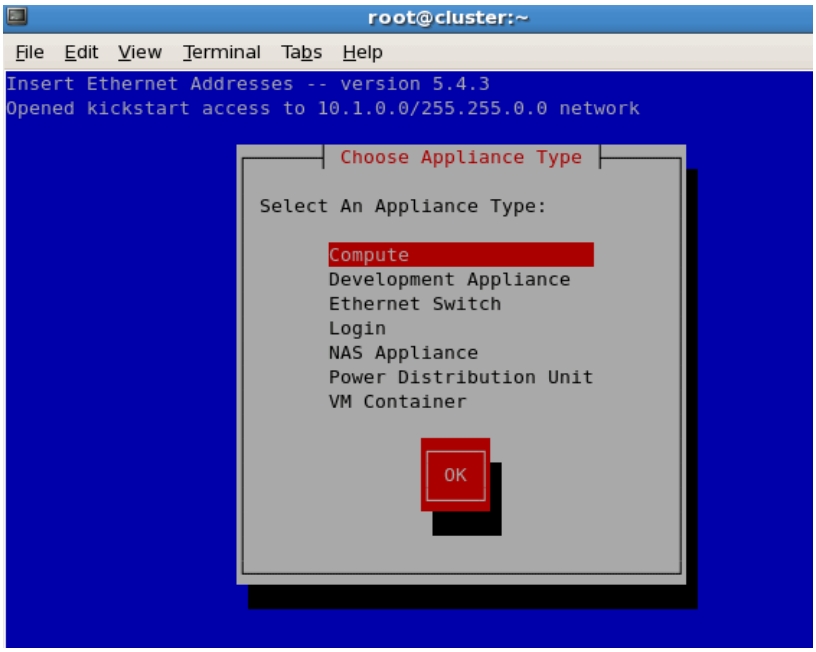
4.5.2 Εγκατάσταση Rocks Cluster στο ¨Node-0¨

Για την εγκατάσταση του Rocks Cluster v5.4.3 viper στον κόμβο όπου έχουμε δημιουργήσει με την ονομασία ¨Node-0¨ θα πρέπει να «φορτώσουμε» το αρχείο ¨.iso¨ όπου είναι το Λειτουργικό μας Σύστημα, με την ίδια διαδικασία όπως έγινε και με το ¨Fronted¨ . Πριν όμως ενεργοποιήσουμε το ¨Node-0¨, θα πρέπει να ανοίξουμε την Γραμμή Εντολών (Command Line) στον υπολογιστή ¨Fronted¨ και να πληκτρολογήσουμε την εντολή ¨instert –ethers¨(Εικόνα 4.31). Αυτή η εντολή μπορεί να χρησιμοποιηθεί για να προσαρμόσει το όνομα (Host Name), IP διευθύνσεις και άλλες παραμέτρους του κόμβου (Node) όπου ανακαλύφθηκε κατά την διάρκεια της εγκατάστασης.



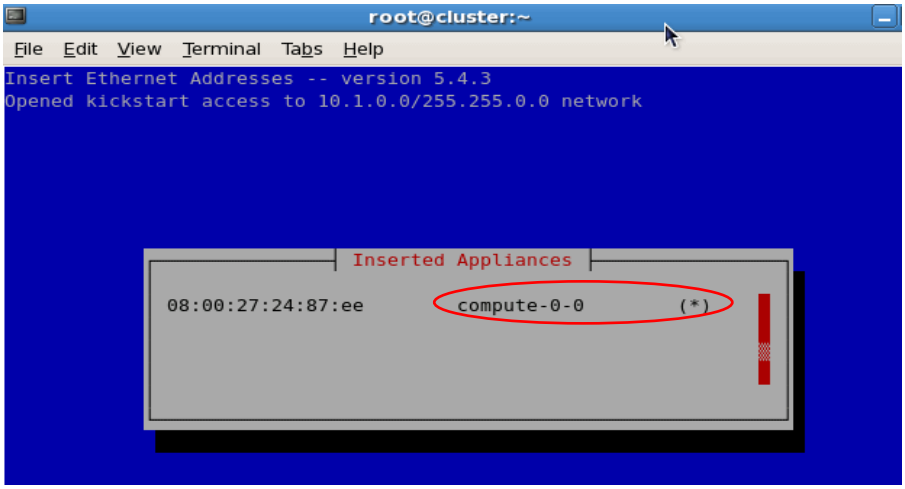
Εικόνα 4.31 Εντολή «Instert –ethers¨ στο Command Line του ¨Fronted¨

Ύστερα από την εκτέλεση της παραπάνω εντολής, εμφανίζεται ένα νέο « παράθυρο» όπως φαίνεται και στην εικόνα 4.32, στο οποίο θα πρέπει να γίνει η επιλογή του μέσου όπου θα γίνει η εκτέλεση-εφαρμογή της παραπάνω εντολής. Στην προκειμένη περίπτωση θα επιλέξουμε το ¨Compute¨ μιας και η συσκευή στην οποία θα εκτελεστεί η εντολή είναι υπολογιστής-κόμβος (Node).



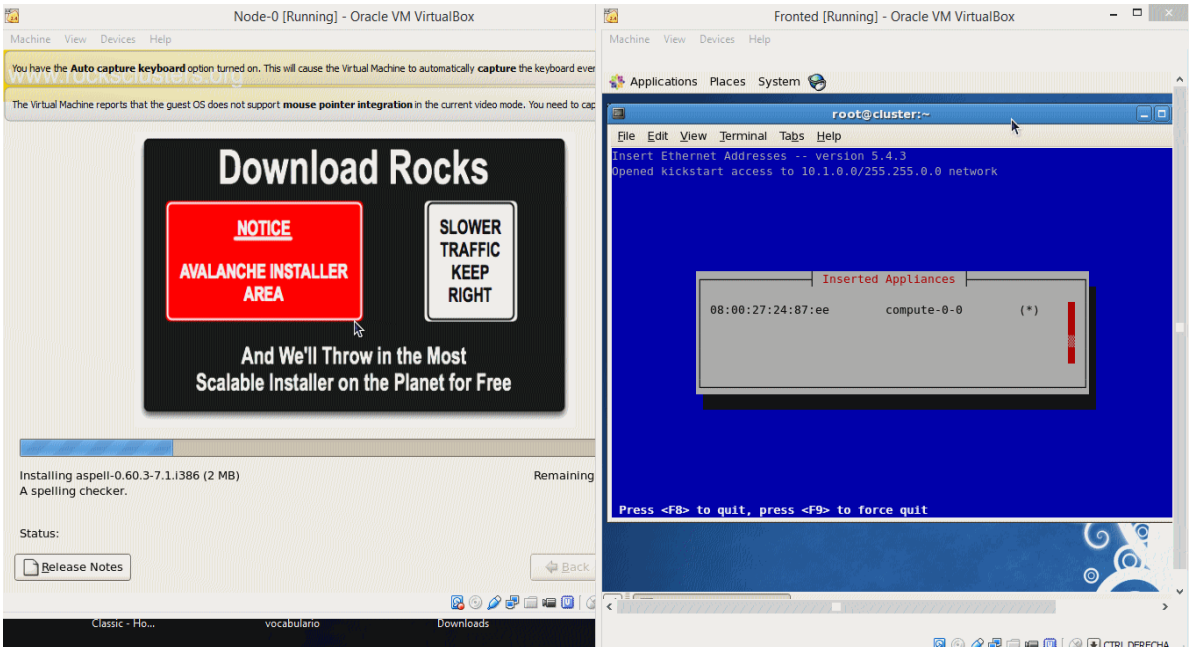
Εικόνα 4.32 Επιλογή ¨compute¨ για την εκτέλεση της εντολής ¨insert –ethers¨

Μετά την επιλογή ¨compute¨ για την εκτέλεση της εντολής ¨insert –ethers¨, πηγαίνουμε στο Virtual Box Manager όπου βρίσκονται οι δύο εικονικοί ηλεκτρονικοί υπολογιστές, (ο ¨Fronted¨ όπου ήδη «τρέχει» και ο ¨Node-0¨) και επιλέγουμε τον κόμβο ¨Node-0¨, όπου και τον ενεργοποιούμε και αρχίζει να φορτώνει το Λειτουργικό Σύστημα (Rocks Cluster v5.4.3 viper) που του είχαμε τοποθετήσει προηγουμένως στο εικονικό CD-DVD ROM. Κατά την εγκατάσταση του Λειτουργικού Συστήματος πραγματοποιείτε επικοινωνία μεταξύ ¨Node-0¨ και ¨Fronted¨, για την προσαρμογή νέων παραμέτρων στο ¨Node-0¨ (Host name, IP, κλπ.). Εάν η επικοινωνία ¨Fronted¨, ¨Node-0¨ είναι επιτυχείς θα πρέπει να πάρουμε το παρακάτω μήνυμα (Εικόνα 4.33).



Εικόνα 4.33 Επιτυχείς πραγματοποίηση επικοινωνίας ¨Fronted¨ με ¨Node-0¨ και προσαρμογή νέων παραμέτρων. Με όνομα Host name ¨compute-0-0¨ για το ¨Node-0¨.

Με αυτόν τον τρόπο εγκαθιστούμε το Λειτουργικό Σύστημα στο ¨Node-0¨ αλλά και παράλληλα πραγματοποιούμε επικοινωνία μεταξύ του ¨Fronted¨ και του ¨Node-0¨ για την προσαρμογή νέων παραμέτρων στο δεύτερο, δίνοντάς του Host name ¨compute-0-0¨ (Εικόνα 4.34). Σε αυτό το σημείο να επισημάνουμε ότι εάν θέλουμε να προσθέσουμε και άλλους κόμβους-nodes στο σύστημά μα, θα ακολουθήσουμε την ίδια διαδικασία και κατά την εκτέλεση της εντολής ¨insert –ethers¨ θα αποδίδονται νέες IP διευθύνσεις στο Internal Network καθώς και Host name, παραδείγματος χάρη το επόμενο θα είναι το ¨compute-0-1¨ και ούτω καθεξής.



Εικόνα 4.34 Εγκατάσταση Rocks Cluster στο ¨Node-0 (στα αριστερά)¨ και επιτυχείς επικοινωνία¨Fronted¨ με¨Node-0¨(στα δεξιά),για την προσαρμογή παραμέτρων.

Ολοκληρώνοντας την εγκατάσταση του Λειτουργικού Συστήματος και όλων των ρυθμίσεων επιτυχώς, έχουμε δημιουργήσει μια συστοιχία cluster τύπου Load Balancer η οποία αποτελείτε από έναν server με την ονομασία ¨Fronted¨ και έναν κόμβο-node με την ονομασία ¨Node-0¨ και με host name ¨compute-0-0¨.

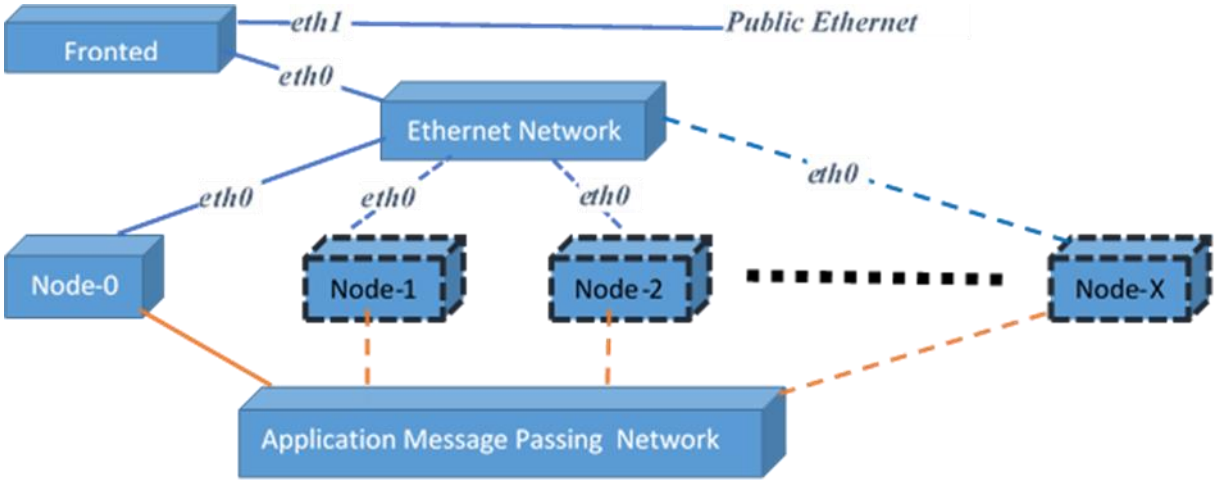
4.5.3 Λειτουργεία του Cluster

Σε ένα εικονικό περιβάλλον (virtual) σαν και αυτό που δημιουργήσαμε το cluster μας, δεν μπορούμε να δούμε καθαρά τα οφέλη μιας συστοιχίας σαν και αυτή, διότι είναι μια προσομοίωση. Έτσι τα δεδομένα μας που θέλουμε να επεξεργαστούμε μπορούν να μοιραστούν από τον ¨Fronted¨ στον κόμβο ¨Node-0¨ ή και σε άλλους κόμβους-nodes αν δημιουργήσουμε στο μέλλον. Με αυτόν τον τρόπο επιτυγχάνουμε καλύτερη αξιοποίηση των διαθέσιμων πόρων

των συστημάτων, μεγαλύτερη ισχύ και ασφάλεια των δεδομένων μας σε περίπτωση που κάποιος κόμβος-node αστοχήσει.

4.5.4 Σχεδιάγραμμα του Cluster

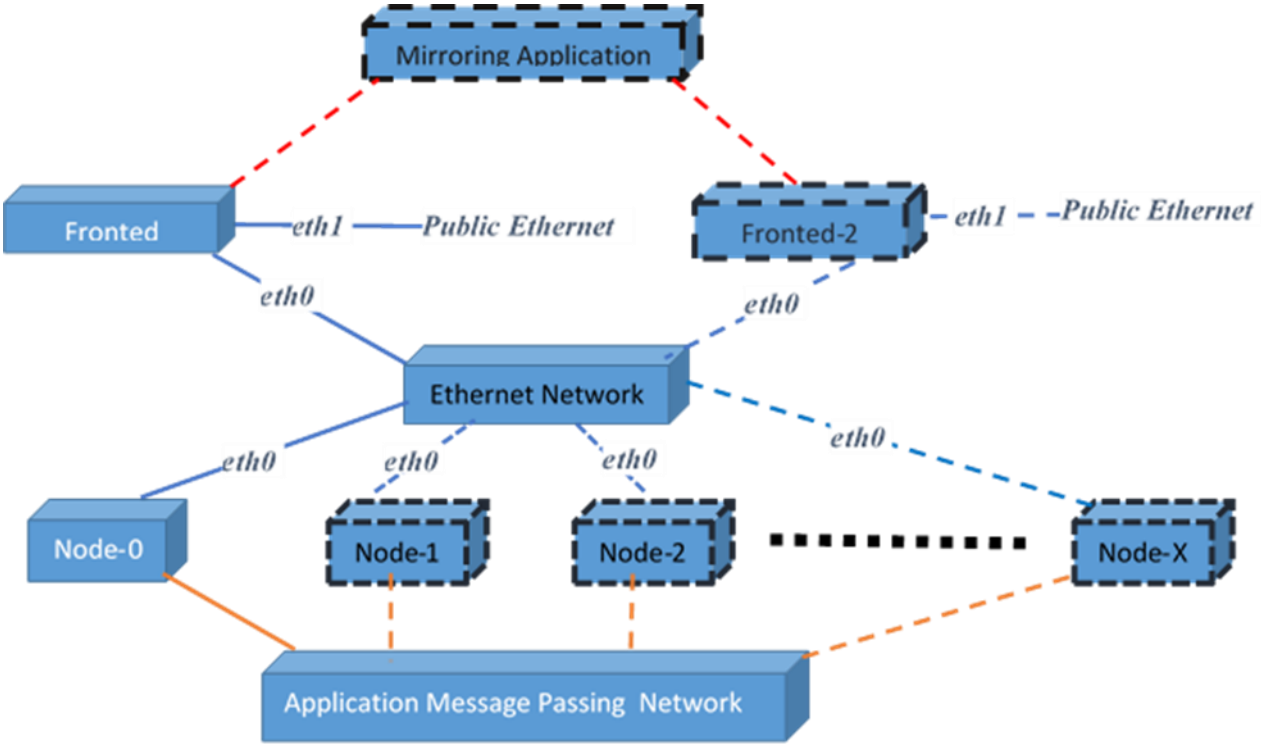
Στο παρακάτω σχεδιάγραμμα μπορούμε να δούμε την δομή του cluster που έχουμε δημιουργήσει (Σχήμα 4.1), καθώς και πιθανές επεκτάσεις της συστοιχίας μας και ως προς τους κόμβους-nodes αλλά και ως προς τους servers (Σχήμα 4.2). Επεκτείνοντας την συστοιχία ως προς τους κόμβους-nodes, τα δεδομένα μας προς επεξεργασία μοιράζονται στους κόμβους-nodes και με αυτόν τον τρόπο η επεξεργασία τους γίνεται ποιο γρήγορα απ΄ ότι σε έναν συνηθισμένο υπολογιστή με τις δυνατότητες ενός απλού κόμβου-node ή ακόμα και server, καθώς γίνεται καταμερισμός της εργασίας και επιστρέφουν το αποτέλεσμα της, στον server (¨Fronted¨). Επίσης εάν ένας κόμβος-node αστοχήσει, δεν επηρεάζει την υπόλοιπη συστοιχία και η επεξεργασία μπορεί να συνεχιστεί στους υπόλοιπους κόμβους-nodes και στον server, καθώς ο κύριος όγκος των δεδομένων δεν αποθηκεύεται συνήθως τοπικά στους κόμβους αλλά στον server της συστοιχίας.



Σχήμα 4.1 Σχεδιάγραμμα του Cluster με πιθανές επεκτάσεις κόμβων-nodes

Όταν γίνεται επέκταση του cluster ως προς τους servers επιτυγχάνουμε μεγαλύτερη αξιοπιστία των δεδομένων μας, καθώς ακόμα και σε περίπτωση αστοχίας του ενός server στην προκειμένη περίπτωση του ¨Fronted¨, έχοντας έναν δεύτερο server για παράδειγμα ¨Fronted-2¨, μπορούμε να ανακτήσουμε όλα τα δεδομένα μας καθώς και να συνεχίσουμε την επεξεργασία τους από το ίδιο σημείο όπου σταμάτησε μετά την αστοχία του πρώτου server ¨Fronted¨. Αυτό επιτυγχάνεται με μία τεχνική αποθήκευσης-backup των δεδομένων μας όπου ονομάζεται mirroring. Με αυτή την τεχνική ότι αποθηκεύτε στον σκληρό δίσκο (ή στους

σκληρούς δίσκους) του πρώτου server ¨Fronted¨, μεταφέρεται και αποθηκεύεται-backup στους σκληρούς δίσκους του δεύτερου server ¨Fronted-2¨ , έτσι διασφαλίζεται η ομαλή και ασφαλής λειτουργία του cluster μας, ιδίως όταν έχουμε να κάνουμε κρίσιμες εφαρμογές και ευαίσθητα δεδομένα.



Σχήμα 4.2 Σχεδιάγραμμα του cluster με πιθανές επεκτάσεις και ως προς τους κόμβους αλλά και ως προς τους servers.

4.5.5 Πρόσθετα υποσυστήματα και χαρακτηριστικά

Η υλοποίηση των παραπάνω στην πράξη, δηλαδή με πραγματικά συστήματα και όχι με εικονικούς ηλεκτρονικούς υπολογιστές και υποσυστήματα (virtual machines), απαιτεί την χρίση κάποιον επιπλέον υποσυστημάτων, όπως είναι ένα switch υψηλής ταχύτητας (τουλάχιστον 1Gbps) για την επικοινωνία των κόμβων μεταξύ τους αλλά και με το ¨Fronted¨, οπότε χρειάζονται καλώδια για την διασύνδεση UTP Cat 5E, που υποστηρίζουν ταχύτητες μεταφοράς μέχρι και 1Gbps. Επίσης ένα KVM switch (Keyboard Video Mouse) για την διασύνδεση όλων των κόμβων αλλά και του ¨Fronted¨ με μία οθόνη, πληκτρολόγιο και ποντίκι, για τον χειρισμό του cluster. Οι θύρες του KVM εξαρτώνται από το πλήθος των κόμβων που θέλουμε να προσθέσουμε καθώς επίσης πρέπει να προβλέπουν και μελλοντικές επεκτάσεις, αυτό μπορεί να επιτευχθεί και με την προσθήκη επιπλέον KVM στο ήδη υπάρχον.

Αναφορές και βιλιογραφία

[1] "Internal Networking". VirtualBox.com Retrieved 2013-07-31.

https://www.virtualbox.org/manual/ch06.html

[2] "Intel Virtualization Technology: Hardware Support for Efficient Processor Virtualization". Intel.com. 2013-08-10. https://en.wikipedia.org/wiki/X86_

virtualization

[3] Wei Huang, Introduction of AMD Advanced Virtual Interrupt Controller, XenSummit 2013. Retrieved 2013-08-11. https://en.wikipedia.org/wiki/X2APIC

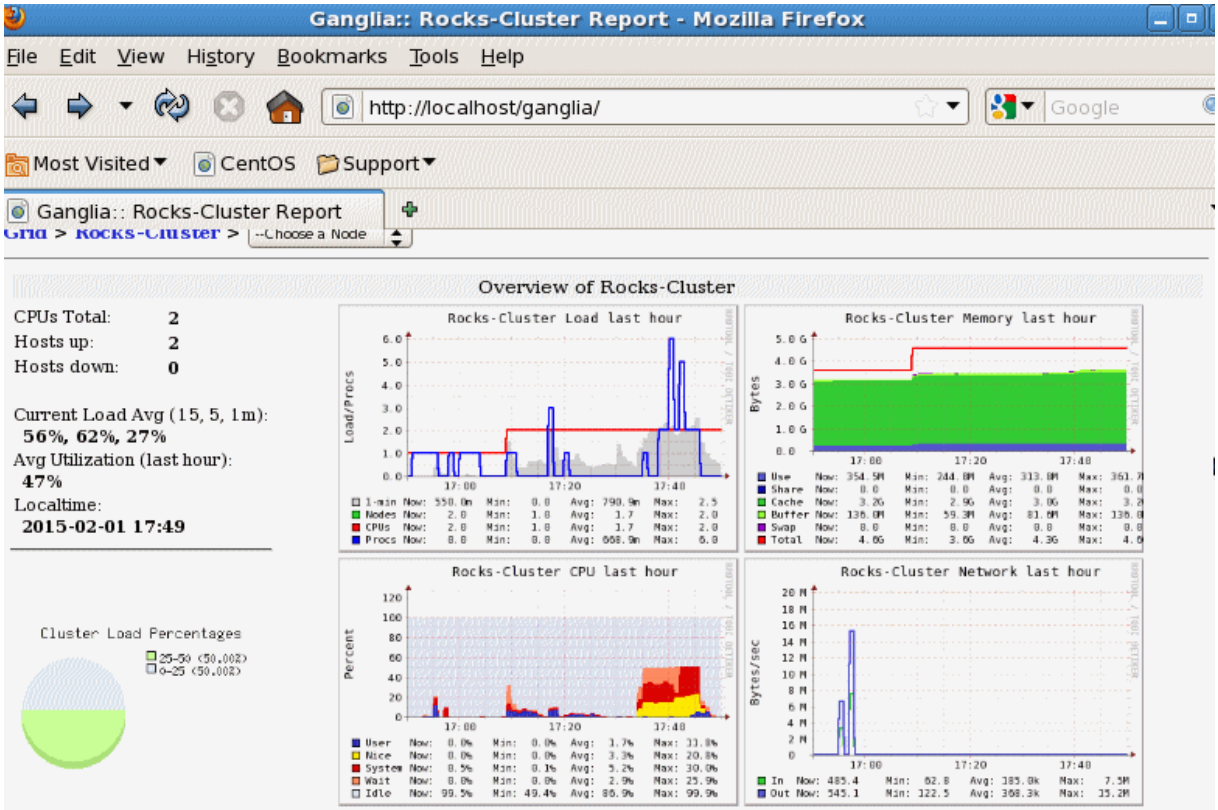
ΚΕΦΑΛΑΙΟ 5

Αποτελέσματα και μετρήσεις

5.1 Προεπισκόπηση του cluster με το Ganglia

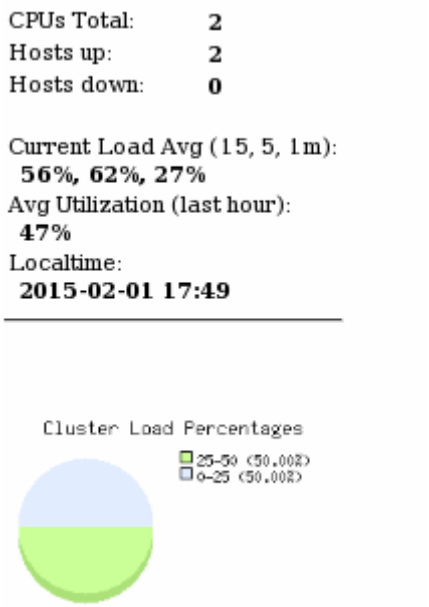
Όπως είχαμε αναφέρει και σε προηγούμενο κεφάλαιο το Ganglia είναι ένα επεκτάσιμο εργαλείο σε που χρησιμοποιείτε σε κατανεμημένα συστήματα (distributed), για υψηλής απόδοσης υπολογιστικά συστήματα όπως είναι τα cluster και τα grids συστήματα. Επιτρέπει στον χρήστη να βλέπει εξ αποστάσεως ζωντανά ή ιστορικά στατιστικά στοιχεία (όπως η μέση τιμή φορτίου της CPU ή του δικτύου) για όλους τους υπολογιστές που παρακολουθούνται. Έτσι κάνοντας χρίση του Ganglia μπορούμε να δούμε στη πράξη το cluster μας να λειτουργεί σε πραγματικό χρόνο, καθώς τα δεδομένα μας ανανεώνονται σε τακτά χρονικά διαστήματα εξ ορισμού ή ακόμα και δυναμικά την στιγμή που εμείς το ζητήσουμε.

Για να δούμε τα αποτελέσματα μέσο του Ganlia θα πρέπει, χρησιμοποιώντας τον φυλλομετρητή, για παράδειγμα ο Firefox και να συνδεθούμε στο http://locaclhost/ganglia/, σε αυτή τη διεύθυνση φιλοξενούνται τα δεδομένα του cluster μας όπως προαναφέραμε. Και η μορφή τους παρουσιάζεται στην παρακάτω εικόνα (Εικόνα 5.1). Στα αριστερά παρουσιάζονται τα συνολικά χαρακτηριστικά του cluster (πλήθος επεξεργαστών, πλήθος υπολογιστών κλπ.) και στα δεξιά μπορούμε να δούμε τον φόρτο του cluster σε συνάρτηση με τον χρόνο, με στατιστική αναπαράσταση.



Εικόνα 5.1 Βλέποντας το cluster σε λειτουργία μέσο του Ganglia

Βλέποντας αναλυτικότερα τα αποτελέσματα στο Ganglia, μπορούμε να δούμε τις διακυμάνσεις κατά την πάροδο του χρόνου ανάλογα με τον φόρτο εργασίας που αναθέτουμε στο cluster, καθώς και αν προσθέτουμε ή αφαιρούμε έναν κόμβο στο cluster μας. Αυτές οι διακυμάνσεις αποτυπώνονται στον επεξεργαστή (CPU), στην μνήμη (RAM) και στην κάρτα δικτύου και αφορούν συνολικά το cluster μας, δηλαδή τα χαρακτηριστικά του ¨Fronted¨ αλλά και του ¨Node-0¨ παρουσιάζονται ως ένας ενιαίος υπολογιστής. Έτσι στο συγκεκριμένο cluster έχουμε δύο επεξεργαστές (CPU) στα 2.4Ghz, 5GB RAM μνήμη καθώς και έναν ενιαίο δίκτυο επικοινωνίας. Όλα αυτά μπορούμε να τα δούμε στα παρακάτω γραφήματα που πάρθηκαν κατά την διάρκεια λειτουργίας του cluster.



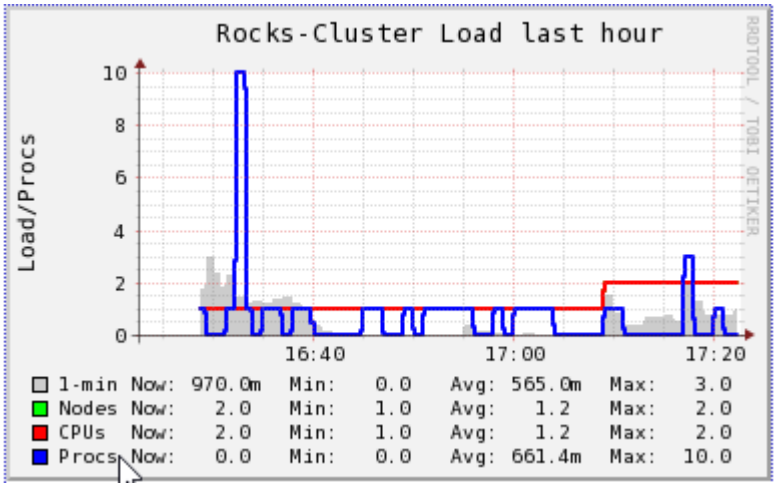
Εικόνα 5.2 Συνολικά χαρακτηριστικά του Cluster

Όπως μπορούμε να δούμε και στην εικόνα 5.2, τα συνολικά χαρακτηριστικά του cluster, έχουμε δυο επεξεργαστές σε λειτουργία, καθώς και δύο ¨hostes up¨ που σημαίνει ότι δύο υπολογιστές από το cluster μας βρίσκονται σε λειτουργία, ο ¨Fronted¨ και ο ¨Node-0¨ και ¨hosts down¨ μηδέν διότι στο συγκεκριμένο cluster δεν υπάρχουν απενεργοποιημένοι υπολογιστές, καθώς υπάρχουν μόνο δύο και βρίσκονται σε λειτουργία. Θα μπορούσε όμως σε μια μελλοντική επέκταση των κόμβων να έχουμε περισσότερους υπολογιστές και κάποιοι να βρίσκονται σε λειτουργία και άλλοι να είναι απενεργοποιημένη ή να δουλεύει το cluster μας στο 100% της απόδοσής του με όλους του κόμβους ενεργοποιημένους, όπως συμβαίνει στο δικό μας cluster. Επίσης μπορούμε να δούμε τον μέσο φόρτο εργασίας στο cluter μας σε συνάρτηση με το χρόνο, όπως είχαμε προαναφέρει σε προηγούμενη παράγραφο. Καθώς και την ώρα και ημερομηνία που πάρθηκαν τα αποτελέσματα. Τέλος παρουσιάζεται με γράφημα σε μορφή «πίτα» ο φόρτος εργασίας που αναλαμβάνει ο κάθε υπολογιστής του cluster μας. Όπως μπορούμε να δούμε στην εικόνα 5.2 ο φόρτος μοιράζεται στους δύο υπολογιστές του

cluster το ποσοστό 50-50 τις εκατό. Η κατανομή της εργασίας ασφαλώς εξαρτάται από το αν οι υπολογιστές είναι ισότιμοι στο cluster, το πλήθος των υπολογιστών που είναι ενεργεί, αλλά και από την δυνατότητα της εφαρμογής να μπορεί να αξιοποιήσει συνολικά τους διαθέσιμους πόρους. Το δικό μας cluster είναι τύπου Κατανομής Φορτίου (Load Balancer) αυτό σημαίνει ότι οι υπολογιστές αντιμετωπίζονται ως ισότιμοι.

5.1.1 Προεπισκόπηση επεξεργαστών του cluster

Όπως βλέπουμε και στην εικόνα 5.3, το Ganglia μπορεί να μας δώσει μια λεπτομερή απεικόνιση του κάθε συστήματος του cluster όπως στην προκειμένη περίπτωση των επεξεργαστών που είναι ενεργεί, των κόμβων-nodes, συμπεριλαμβανομένου και του ¨Fronted¨ καθώς θεωρείτε ισότιμος με τα υπόλοιπα nodes και τέλος βλέπουμε και τον φόρτο εργασίας. Με κόκκινο χρώμα απεικονίζονται οι διαθέσιμοι επεξεργαστές στο cluster μας και με μπλε ο φόρτος εργασίας (Load/Procs). Με πράσινο χρώμα είναι τα διαθέσιμα nodes αλλά δεν απεικονίζονται στο γράφημα.

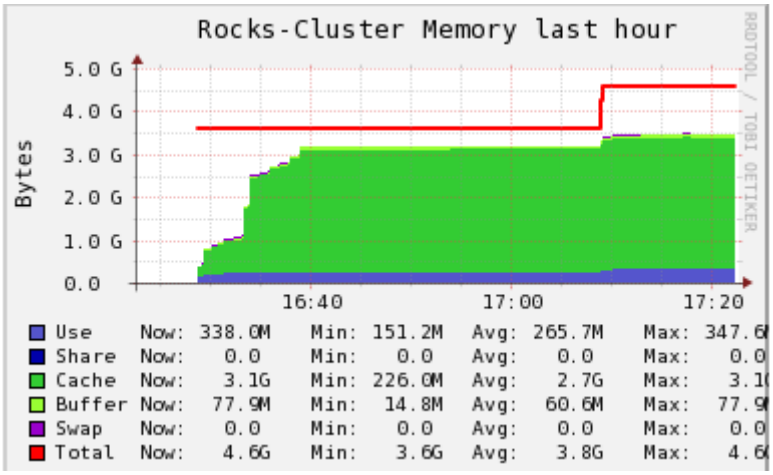


Εικόνα 5.3 Απεικόνιση του φόρτου επεξεργασίας, των nodes , καθώς και των διαθέσιμων επεξεργαστών (CPU΄s), σε συνάρτηση με τον χρόνο.

Αναλυτικότερα όπως βλέπουμε στο γράφημα (Εικόνα 5.3), την γραμμή των επεξεργαστών (κόκκινη) μέχρι την χρονική στιγμή «17:09» έχουμε διαθέσιμο μόνο έναν επεξεργαστή, άρα και μόνο έναν υπολογιστή ενεργό, διότι όπως προαναφέραμε κάθε υπολογιστής στο σύστημά μας διαθέτει έναν επεξεργαστή με έναν πυρήνα. Στην χρονική «17:10» παρατηρείται μία μετατόπιση της κόκκινης γραμμής από την θέση «1» στην θέση «2» του κάθετου άξονα που εκεί απεικονίζοντα οι διαθέσιμοι επεξεργαστές, οπότε έχουμε δύο CPU΄s ενεργές, άρα και δύο υπολογιστές ενεργούς, όπου ο δεύτερος τέθηκε σε λειτουργία την χρονική στιγμή «17:10».

5.1.2 Προεπισκόπηση μνήμης RAM του cluster

Το Ganglia επίσης μας δίνει την δυνατότητα να δούμε αναλυτικότερα την λειτουργία της μνήμης στο cluster μας. Όπως βλέπουμε και από το παρακάτω γράφημα (εικόνα 5.4), μας δίνει πληροφορίες για την χρήσει της μνήμης (μπλε χρώμα), την μνήμη που μοιράζεται μεταξύ των κόμβων (σκούρο μπλε), το cash (πράσινο) και το buffer της μνήμης (ανοιχτό πράσινο) και τέλος την ποσότητα της διαθέσιμης μνήμης (κόκκινο).

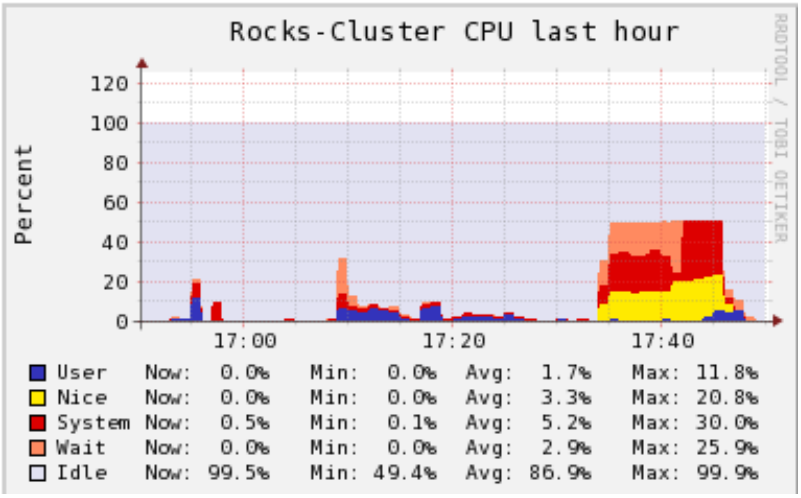


Εικόνα 5.4 Απεικόνιση μνήμης RAM του cluster.

Όπως μπορούμε να δούμε στο γράφημα της εικόνας 5.4 , έως την χρονική στιγμή «17:09» η διαθέσιμη ποσότητα μνήμης είναι περίπου 4GB όσα δηλαδή είχαμε ορίσει στον υπολογιστή ¨Fronted¨. Την χρονική στιγμή «17:10» παρατηρούμε μετατόπιση της κόκκινης γραμμής ως προς τον κάθετο άξονα, όπου απεικονίζεται το ποσό της μνήμης RAM και από 4Gb μετατοπίζεται στα 5GB. Αυτό συμβαίνει διότι την συγκεκριμένη χρονική στιγμή συνδέθηκε στο cluster ο κόμβος-node ¨Node-0¨, όπου στα τεχνικά χαρακτηριστικά του διαθέτει 1GB RAM, άρα με την ενεργοποίησή του προστίθεται στα ήδη 4GB του υπολογιστή ¨Fronted¨.

5.1.3 Χρήση CPU

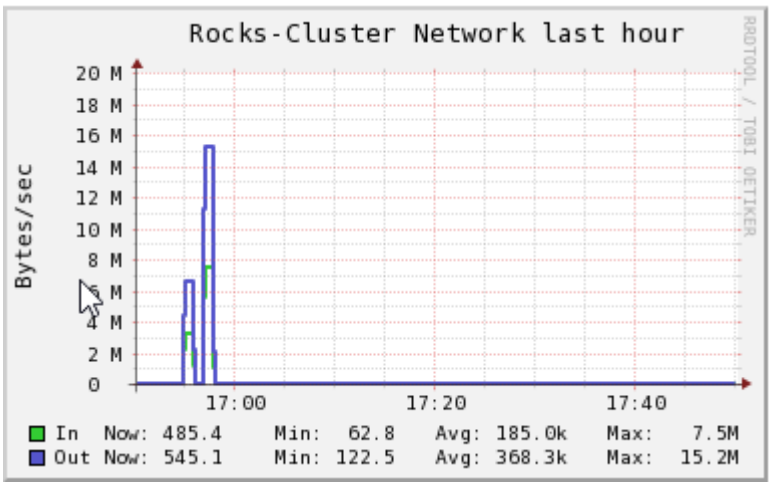
Μέσο του Ganglia μπορούμε να δούμε ακόμα μια γραφική παράσταση, που αφορά λεπτομερώς την χρήση του/ων επεξεργαστή/στων, όπως ανενεργός, σε κατάσταση αναμονής, χρήστης, κλπ. Όπως παρουσιάζεται στο γράφημα της εικόνας 5.5.



Εικόνα 5.5 Αναλυτική χρήση CPU

5.1.4 Κίνηση δικτύου δεδομένων

Ένα ακόμα δεδομένο που μας δίνει το Ganglia, είναι μια γραφική αναπαράσταση της κίνησης των δεδομένων στο δίκτυο του cluster. Η κίνηση στο δίκτυο απεικονίζεται με πράσινο χρώμα για τα δεδομένα που εισάγονται και με μπλε τα δεδομένα που εξέρχονται από αυτό. Η μέτρησή τους γίνεται σε ¨Bytes/sec¨ και ο κάθετος άξονας απεικονίζει των όγκο των δεδομένων σε Megabytes σε συνάρτηση πάντα με τον οριζόντιο άξονα του χρόνου (Εικόνα 5.6).

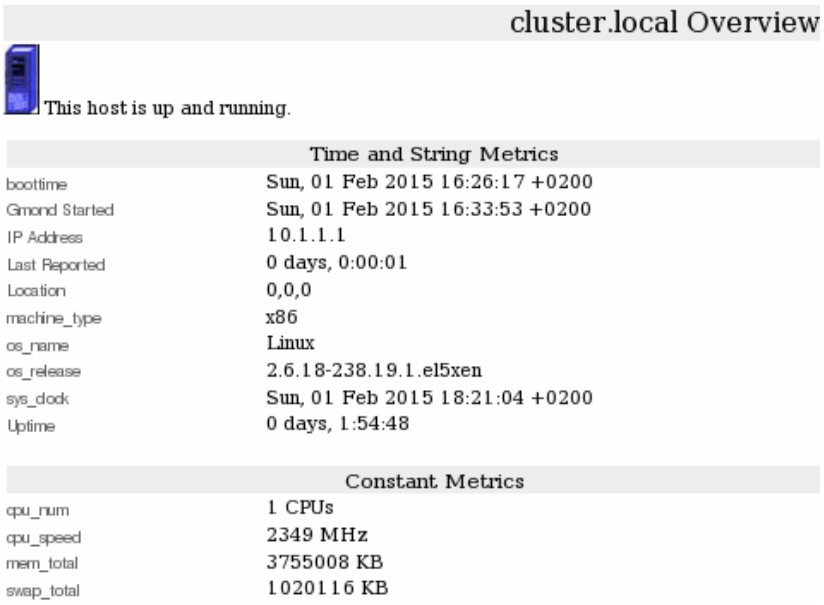


Εικόνα 5.6 Χρήση δικτύου cluster

5.2 Χαρακτηριστικά των Hosts

Το Ganlia μας δίνει ακόμα ποιο λεπτομερή εικόνα για την λειτουργία του cluster μας, δίνοντάς μας ξεχωριστά δεδομένα για τον κάθε υπολογιστή που είναι συνδεδεμένος με το cluster μας, καθώς επίσης και για το κάθε υποσύστημά τους ξεχωριστά, όπως γίνεται και στην περίπτωση των συνολικών στοιχείων του cluster . Όπως για παράδειγμα ο ¨Fronted¨ με host-

name ¨cluster¨ και ο ¨Node-0¨ με host-name ¨compute-0-0¨, που τα χαρακτηριστικά τους αποτυπώνονται στις παρακάτω εικόνες (Εικόνα 5.7)(Εικόνα 5.8).



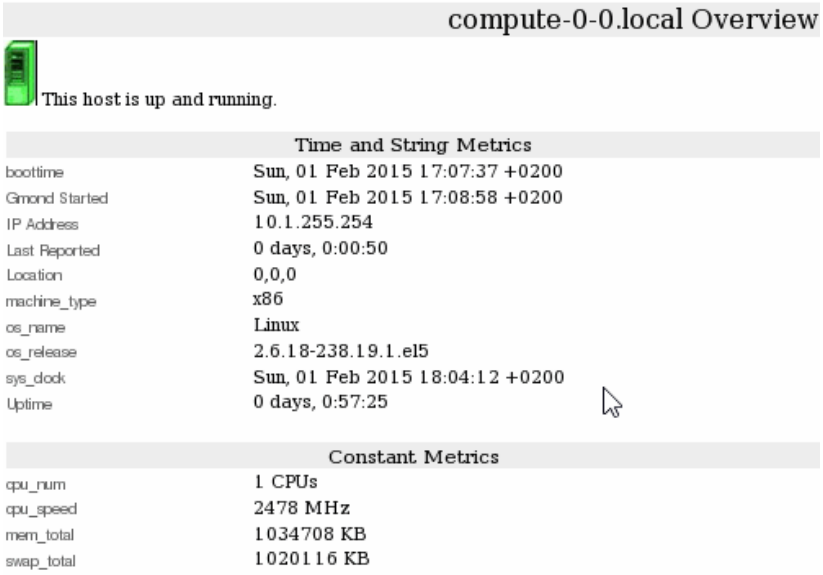
cluster.local Overview

This host is up and running.

Time and String Metrics	
boottime	Sun, 01 Feb 2015 16:26:17 +0200
Gmond Started	Sun, 01 Feb 2015 16:33:53 +0200
IP Address	10.1.1.1
Last Reported	0 days, 0:00:01
Location	0,0,0
machine_type	x86
os_name	Linux
os_release	2.6.18-238.19.1.el5xen
sys_clock	Sun, 01 Feb 2015 18:21:04 +0200
Uptime	0 days, 1:54:48

Constant Metrics	
cpu_num	1 CPUs
cpu_speed	2349 MHz
mem_total	3755008 KB
swap_total	1020116 KB

Εικόνα 5.7 Χαρακτηριστικά ¨Fronted¨ με host-name ¨cluster¨



compute-0-0.local Overview

This host is up and running.

Time and String Metrics	
boottime	Sun, 01 Feb 2015 17:07:37 +0200
Gmond Started	Sun, 01 Feb 2015 17:08:58 +0200
IP Address	10.1.255.254
Last Reported	0 days, 0:00:50
Location	0,0,0
machine_type	x86
os_name	Linux
os_release	2.6.18-238.19.1.el5
sys_clock	Sun, 01 Feb 2015 18:04:12 +0200
Uptime	0 days, 0:57:25

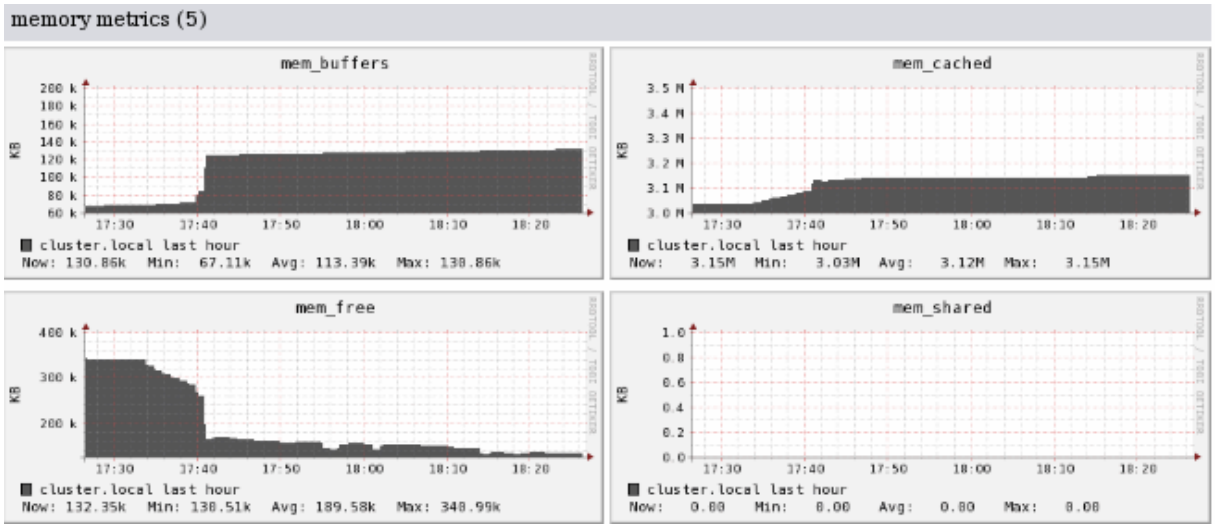
Constant Metrics	
cpu_num	1 CPUs
cpu_speed	2478 MHz
mem_total	1034708 KB
swap_total	1020116 KB

Εικόνα 5.8 Χαρακτηριστικά ¨Node-0¨ με host-name ¨compute-0-0¨

5.2.1 Αναλυτικότερες μετρήσεις των Hosts

Όπως προαναφέραμε υπάρχουν για το κάθε υποσύστημα του κάθε host ξεχωριστά δεδομένα, που καταγράφονται με την βοήθεια του Ganglia και αποτυπώνονται σε γραφικές παραστάσεις, το καθένα host ξεχωριστά (χρήση CPU, χρήση μνήμης, διεργασίες στη CPU και αριθμός CPU, κίνηση δεδομένων στην κάρτα δικτύου του συστήματος). Επιπλέον υπάρχουν

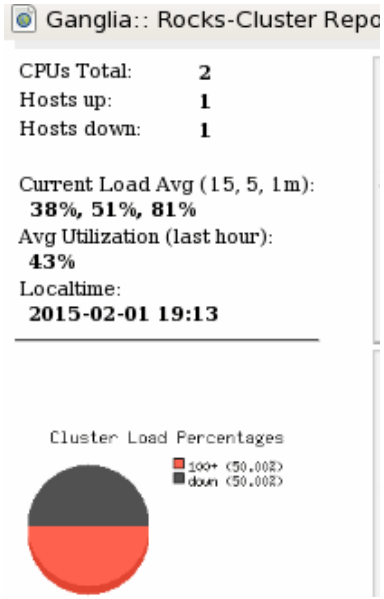
ακόμα ποιο αναλυτικές μετρήσεις για το κάθε από τα παραπάνω υποσυστήματα των hosts. Ένα χαρακτηριστικό παράδειγμα μιας τέτοιας μέτρησης φαίνεται στην εικόνα 5.9 όπου παρουσιάζει αναλυτικά η λειτουργία της μνήμης RAM του¨Fronted¨ .



Εικόνα 5.9 Αναλυτικές μετρήσεις μνήμης ¨Fronted¨

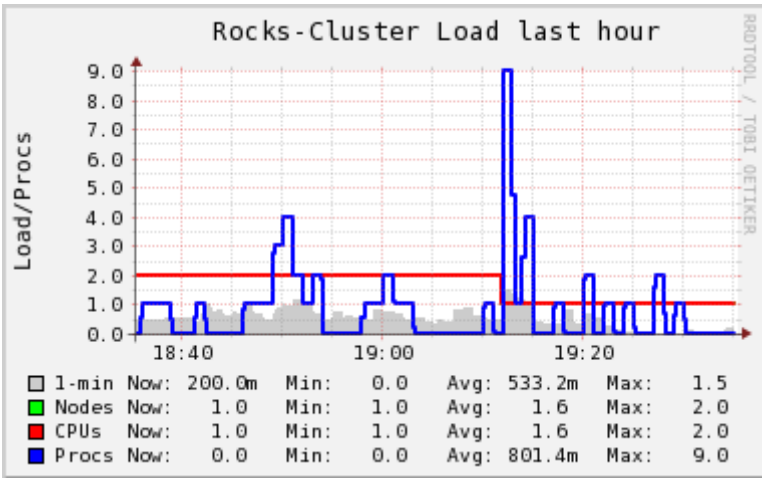
5.3 Απενεργοποίηση ή αστοχία κόμβου-node

Τέλος σε περίπτωση που κάποιος κόμβος-node στο cluster τεθεί εκτός λειτουργίας, είτε οικειοθελώς είτε από αστοχία, τότε μπορούμε αμέσως να το καταλάβουμε και να εντοπίσουμε τον κόμβο μέσο του γενικού συστήματος παρακολούθησης του cluster. Για παράδειγμα ενώ λειτουργούσαν και τα δυο hosts, απενεργοποιήσαμε τον ¨compute-0-0¨ που αντιστοιχεί στο ¨Node-0¨. Στην στατιστική « πίτα» απεικονίζεται με μαύρο χρώμα ο απενεργοποιημένος, ¨host down 1¨.(Εικόνα 5.10)



Εικόνα 5.10 Απενεργοποίηση του host ¨compute-0-0¨ (Node-0).

Σε αντίθετη περίπτωση απενεργοποιήσεις ή αστοχίας του ¨Fonted¨ θα είχαμε κατάρευσει όλου του cluster ακόμα και αν σε ένα τέτοιο σύστημα οι hosts είναι ισότιμοι. Αυτό θα συνέβαινε διότι η εγκατάσταση του Λειτουργικού Συστήματος έχει γίνει στον σκληρό δίσκο του ¨Fronted¨ καθώς και όλες οι ρυθμίσεις του cluster βρίσκονται τοπικά αποθηκευμένες σε αυτόν. Αυτό είναι και ένα από τα κύρια μειονεκτήματα του cluster μας.



Εικόνα 5.11 Από «δύο» σε «ένα», ο αριθμός των CPU΄s μετά την απενεργοποίηση του host ¨compute-0-0¨

5.4 Εφαρμογή SuperPi

Το SuperPi είναι ένα πρόγραμμα υπολογιστή που υπολογίζει το π σε ένα συγκεκριμένο αριθμό των ψηφίων μετά το δεκαδικό σημείο, μέχρι το ανώτατο όριο των 32 εκατομμυρίων και χρησιμοποιεί τον αλγόριθμο Gauss-Legendre.

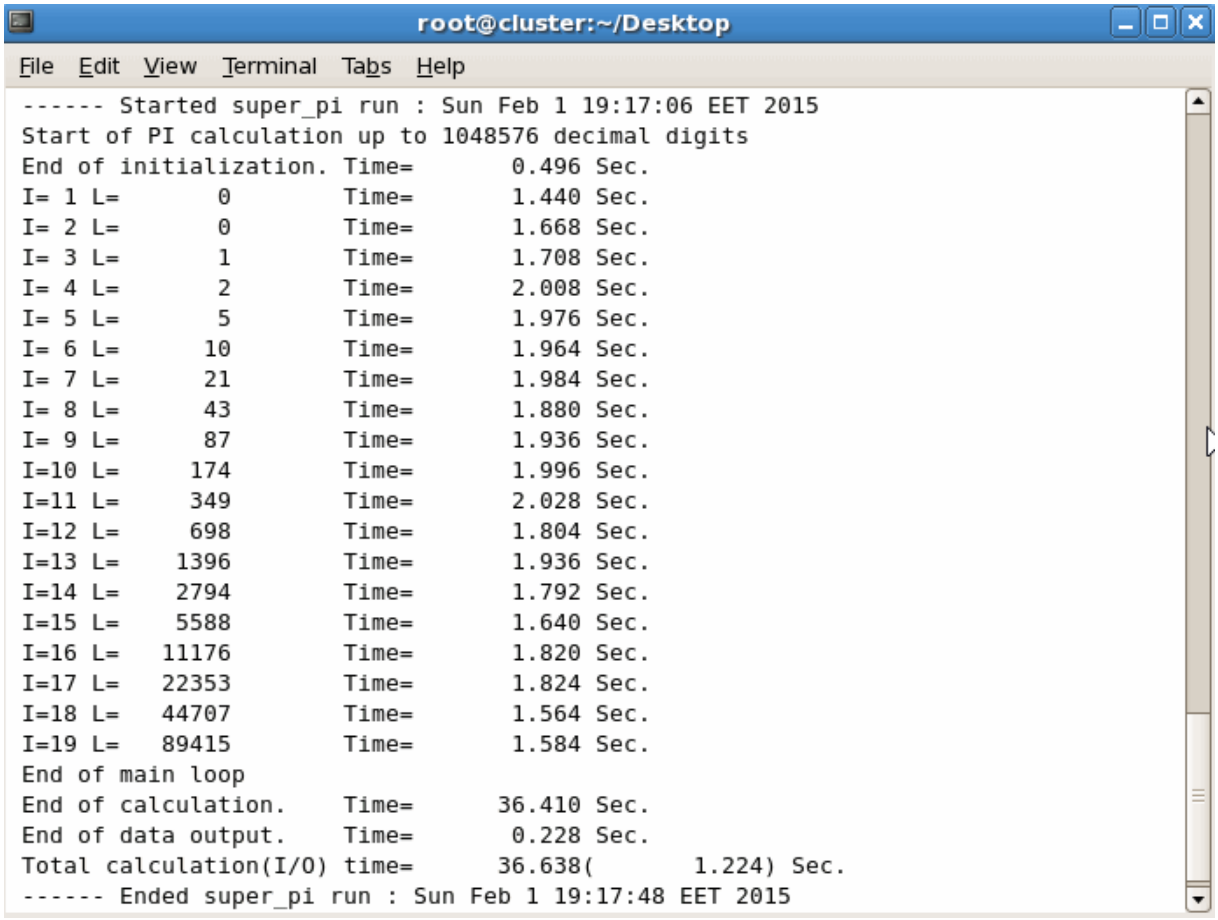
Γενικός τύπος:

$$\pi = 4 \cdot \sum_{i=1}^{\infty} (-1)^{i+1} \frac{1}{2i-1} \qquad (1)$$

Το SuperPi αναπτύχθηκε για συστήματα Windows, αλλά τα τελευταία χρόνια έχουν αναπτυχθεί διάφορες εκδόσεις του προγράμματος για να τρέχει και σε Λειτουργικά Συστήματα Linux. Το SuperPi είναι ένα αξιόπιστο benchmark που πολλοί χρήστες, κυρίως overclockers, χρησιμοποιούν για να μετρήσουν τις επιδώσεις των συστημάτων τους. Μία από αυτές τις εκδώσεις χρησιμοποιούμε και στο cluster μας, συγκεκριμένα την ¨SuperPi mod 1.5¨, για να μπορέσουμε να μετρήσουμε την απόδοση του cluster.

5.4.1 Μετρήσεις με το SuperPi

Για να δούμε την απόδοση του συστήματός μας θα πρέπει να εκτελέσουμε το SuperPi πρώτα για τον έναν host και ύστερα να ενεργοποιήσουμε και τον δεύτερο, ώστε να δούμε συνολικά τα αποτελέσματα και να μπορούμε να κάνουμε μια σύγκριση. Πριν εκελέσουμε το πρόγραμμα θα πρέπει να το ρυθμίσουμε πόσα ψηφία θα υπολογίσει μετά το δεκαδικό μέρος του π. Επίσης έχει θεσπιστεί να εκτελείτε συνήθως για 2^{20} , δηλαδή για 1.048.576 ψηφία. Στην εικόνα 5.12 μπορούμε να δούμε τα αποτελέσματα της εκτέλεσης του SuperPi σε έναν μόνο host και συγκεκριμένα στον ¨Fronted¨. Το αποτέλεσμα που λαμβάνουμε είναι για 2^{20} ψηφία είναι, 36.410 δευτερόλεπτα.



Εικόνα 5.12 Εκτέλεση SuperPi για 2^{20} ψηφία, με έναν host ενεργό και μία CPU.

Εφόσον εκτελέσαμε το πρόγραμμα για έναν host με μία CPU, ύστερα θα πρέπει να θέσουμε σε λειτουργία και τον δεύτερο host (Node-0) και να το ξανά εκτελέσουμε. Όπως μπορούμε να δούμε από τις μετρήσεις στην εικόνα 5.13, με δύο ενεργά host και δύο CPU, ο υπολογισμός για 2^{20} ψηφία εκτελέστηκε σε χρόνο 32.982 δευτερόλεπτα.

```
------ Started super_pi run : Sun Feb 1 19:07:15 EET 2015
Start of PI calculation up to 1048576 decimal digits
End of initialization. Time=       0.496 Sec.
I= 1 L=       0        Time=       1.460 Sec.
I= 2 L=       0        Time=       1.660 Sec.
I= 3 L=       1        Time=       1.672 Sec.
I= 4 L=       2        Time=       1.680 Sec.
I= 5 L=       5        Time=       1.660 Sec.
I= 6 L=      10        Time=       1.668 Sec.
I= 7 L=      21        Time=       1.640 Sec.
I= 8 L=      43        Time=       1.664 Sec.
I= 9 L=      87        Time=       1.692 Sec.
I=10 L=     174        Time=       1.700 Sec.
I=11 L=     349        Time=       1.688 Sec.
I=12 L=     698        Time=       1.696 Sec.
I=13 L=    1396        Time=       1.744 Sec.
I=14 L=    2794        Time=       1.680 Sec.
I=15 L=    5588        Time=       1.648 Sec.
I=16 L=   11176        Time=       1.656 Sec.
I=17 L=   22353        Time=       1.600 Sec.
I=18 L=   44707        Time=       1.568 Sec.
I=19 L=   89415        Time=       1.440 Sec.
End of main loop
End of calculation.    Time=      32.982 Sec.
End of data output.    Time=       0.204 Sec.
Total calculation(I/O) time=      33.186(       1.236) Sec.
------ Ended super_pi run : Sun Feb 1 19:07:49 EET 2015
```

Εικόνα 5.13 Εκτέλεση SuperPi με δύο host ενεργά και δύο CPU.

5.4.2 Συνολικά αποτελέσματα SuperPi

Μετά τις παραπάνω μετρήσεις με ένα και δύο host computers, συγκεντρώσαμε τα αποτελέσματά τους σε έναν πίνακα για να μπορέσουμε να κάνουμε την απαραίτητη σύγκριση. Όπως βλέπουμε και από τον πίνακα 1, ο συνολικός χρόνος που χρειάζεται για να εκτελεστεί ο υπολογισμός 2^{20} ψηφίων μετά το δεκαδικό μέρος του π, με χρήση μόνο του ενός host (Fronted), χρειάζεται 36.410 δευτερόλεπτα για να ολοκληρωθεί. Σε αντίθεση με δύο host, δηλαδή με την προσθήκη και του ¨Node-0¨, ο ίδιος υπολογισμός χρειάζεται 32.982 δευτερόλεπτα για να ολοκληρωθεί.

Έχοντας τα αποτελέσματα αυτά μπορούμε να συμπεράνουμε ότι, λειτουργώντας το cluster μας με δύο host είναι σαφές ποιο γρήγορο απ' ότι μόνο με τον ¨Fronted¨. Η διαφορά των δύο μετρήσεων είναι στα 3,428 δευτερόλεπτα υπέρ του cluster.

Πλήθος Κόμβων (Hosts)	Πλήθος CPU	Χρόνος Εκτέλεσης
1	1	36.410 sec
2	2	32.982 sec

Πίνακας 5.1 Εκτέλεση του SuperPi

5.4.3 Παρατηρήσεις

Αυτή η διαφορά αν και δεν φαίνεται πολύ μεγάλη, είναι ιδιαίτερα σημαντική διότι μιλάμε για μόνο δύο ηλεκτρονικούς υπολογιστές, κάτι που αποτελεί το μίνιμουμ για έναν cluster. Σε αυτό θα πρέπει να συνυπολογίσουμε και τις δυνατότητες των hosts καθώς διαθέτουν μόνο έναν επεξεργαστή (CPU) ο καθένας, ενώ σε άλλη περίπτωση θα μπορούσε ένας host να διαθέτει δύο ή και περισσότερους επεξεργαστές. Ένας ακόμα σημαντικός παράγοντας που δεν θα πρέπει να παραλείψουμε είναι ότι, τα hosts δηλαδή οι ηλεκτρονικοί υπολογιστές που κάνουμε τις μετρήσεις μας είναι εικονικοί και δεν έχουν φυσική υπόσταση, άρα μοιράζονται τους ίδιους πόρους (CPU, RAM, σκληρός δίσκος, κάρτα δικτύου, κλπ) του φυσικού συστήματος όπου «τρέχουν», δηλαδή το laptop Sony Vaio SVE1513C1EW. Τέλος οι μετρήσεις μπορούν να επηρεαστούν και από εξωγενείς παράγοντες, όπως λειτουργίες που εκτελούνται στο παρασκήνιο-background (προγράμματα, antivirus, updates, κλπ) του φυσικού υπολογιστή. Γι΄ αυτό σε συνεχόμενες μετρήσεις που έγιναν εντοπίστηκε διακύμανση των αποτελεσμάτων. Παρόλο αυτά δεν αναιρεί το συμπέρασμα ότι, λειτουργώντας περισσότεροι του ενός ηλεκτρονικοί υπολογιστές μαζί στη μορφή ενός cluster, επιτυγχάνουν μεγαλύτερη απόδοση.

Αναφορές και Βιβλιογραφία

- Martinović, G.; Balen, J.; Rimac-Drlje, S. (2010), "Impact of the host operating systems on virtual machine performance", *2010 Proceedings of the 33rd International Convention MIPRO*, IEEE, pp. 613–618. 2014-05-07. https://www.researchgate.net/publication /224162915_Impact_of_the_host_operating_systems_on_virtual_machine_performance.

- Sanchez, Ernesto; Squillero, Giovanni; Tonda, Alberto (2011), "Evolutionary Failing-test Generation for Modern Microprocessors" (PDF), *Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation (GECCO '11)*, New York, NY, USA: ACM, pp. 225–226, doi:10.1145/2001858.2001985, ISBN 978-1-4503-0690-4. 2014-05-10 https://www.researchgate.net/publication/220742782_ Evolutionary_ failing-test_generation_for_modern_microprocessors.

- *Round 2... 10 Trillion Digits of Pi*, numberworld.org http://phys.org/news/2011-10-pi-enthusiast-ten-trillionth-digit.html 2014-05-10

ΚΕΦΑΛΑΙΟ 6

Συμπεράσματα και προτάσεις περεταίρω μελέτης

6.1 Συμπεράσματα

Η συστοιχία είναι κατά βάση μία συστοιχία εξισορρόπησης φορτίου (Load Balancing Cluster). Αυτό σημαίνει ότι κάθε φορά που ξεκινά μία νέα εφαρμογή να εκτελείται, το σύστημα μετακινεί διεργασίες μεταξύ των κόμβων της συστοιχίας έτσι ώστε το φορτίο των δύο διαθέσιμων κόμβων να είναι περίπου το ίδιο (στην προκειμένη περίπτωση 50-50). Η διαδικασία αυτή γίνεται αυτόματα και χωρίς παρέμβαση του χρήστη, ενώ ούτε οι διεργασίες καταλαβαίνουν τη διαφορά, αφού η συστοιχία είναι διαφανής σε αυτές. Ωστόσο η συστοιχία μπορεί να χρησιμοποιηθεί και για την ταχύτερη εκτέλεση προγραμμάτων (και κυρίως αυτών που αφορούν αριθμητικούς υπολογισμούς), λειτουργώντας έτσι σαν μια συστοιχία υψηλής απόδοσης (HPC-High Performance Computing). Απόδειξη αυτού είναι η εκτέλεση του SuperPi και των αποτελεσμάτων του που είδαμε στο Κεφάλαιο 5, όπου η εκτέλεση απαιτεί 3.428 δευτερόλεπτα λιγότερα από το να «έτρεχε» μόνο στον έναν υπολογιστή. Αυτό ασφαλώς θα ήταν περισσότερο εμφανές σε ένα «φυσικό» cluster και όχι τόσο σε ένα εικονικό σαν το δικό μας.

6.1.1 Πλεονεκτήματα

Τα πλεονεκτήματα ενός cluster σαν του δικού μας είναι:

1) Υπολογιστική Ισχύς: Η δυνατότητα να αξιοποιήσουμε την συνδυασμένη υπολογιστική ενός cluster σαν του δικού μας μπορεί σε πολλές περιπτώσεις να αποδειχθεί αποτελεσματικότερη, σε σχέση με το κόστος, ενός υπερυπολογιστή, παρόμοιων δυνατοτήτων. Με τον τρόπο αυτό μια επιχείρηση ή ένας οργανισμός μπορεί να αξιοποιήσει καλύτερα τον εξοπλισμό που έχουν διαθέσιμο χωρίς σημαντικά παραπάνω έξοδα.

2) Μείωση Κόστους: Στις μέρες μας το κόστος των προσωπικών υπολογιστών έχει μειωθεί σημαντικά, από την άλλη η μείωση αυτή συνοδεύεται από εκρηκτική αύξηση των επιδόσεων και της υπολογιστικής τους ισχύς. Ένας σημερινός προσωπικός υπολογιστής μεσαίων δυνατοτήτων, όπως είναι οι υπολογιστές που χρησιμοποιήσαμε στο παράδειγμά μας για την κατασκευή του cluster, είναι πολλές φορές ποιο ισχυρός απ' τους πρώτους υπερυπολογιστές (Super Computers).

3) Επεκτασιμότητα: Ένα από τα πολύ μεγάλα πλεονεκτήματα ενός cluster σαν του δικού μας είναι η επεκτασιμότητα, καθώς μπορούμε πάρα πολύ εύκολα και με χαμηλό κόστος να επεκτείνουμε το cluster μας, προσθέτοντας και άλλους κόμβους ίδιων ή ακόμα και διαφορετικών προδιαγραφών. Όπως είδαμε και στο cluster μας ο ¨Fronted¨ και ο ¨Node-0¨ έχουν τελείως διαφορετικά χαρακτηριστικά.

6.1.2 Μειονεκτήματα

Το μεγαλύτερο μειονέκτημα που εντοπίζεται σε ένα cluster σαν αυτό που κατασκευάσαμε είναι:

1) Διαθεσιμότητα: Αν και σε ένα cluster τύπου Load Balancing όλοι οι κόμβοι είναι ισότιμοι, αυτό ισχύει μόνο για την εκτέλεση των προγραμμάτων σε αυτούς, εάν ένας ¨server¨ τεθεί εκτός λειτουργίας, όπως είναι ο ¨Fronted¨ για την δική μας συστοιχία, τότε όλο το σύστημα καταρρέει. Αυτό συμβαίνει διότι στον σε έναν κόμβο του συστήματος όπως είναι ο ¨Fronted¨ έχει γίνει εγκατάσταση του Λειτουργικού Συστήματος στο σκληρό δίσκο κάτι που στους υπόλοιπους κόμβους-nodes δεν είναι απαραίτητο, καθώς και όλες οι ρυθμίσεις για τον εντοπισμό και την επικοινωνία των κόμβων μεταξύ τους βρίσκονται εγκατεστημένες τοπικά στον κόμβο ¨Fronted¨. Τέλος ο κόμβος ¨Fronted¨ είναι ο μοναδικός που επικοινωνεί με τον «έξω κόσμο» ή αλλιώς internet. Όλα αυτά τα παραπάνω καθιστούν αυτόν τον κόμβο-server πολύ σημαντικό για το cluster και τυχών δυσλειτουργία του θα σήμαινε και κατάρρευση όλου του συστήματος. Όπως είχαμε εξηγήσει σε προηγούμενο κεφάλαιο αυτό το πρόβλημα μπορεί να αντιμετωπιστεί προσθέτοντας έναν δεύτερο server με λειτουργία mirroring, κάτι τέτοιο βέβαια θα αύξανε το κόστος του cluster αλλά θα μας διασφάλιζε σε περίπτωση δυσλειτουργίας του πρώτου server. Σε αντίθεση με το τι συμβαίνει στην περίπτωση αστοχίας του κόμβου-server, όταν τεθεί εκτός λειτουργίας ένας κόμβος-node ή περισσότεροι, το cluster μας μπορεί να εξακολουθεί να λειτουργεί χωρίς κανένα πρόβλημα.

Παρατήρηση: Ένα ακόμα πρόβλημα που όμως εντοπίζεται σε εικονικά και μόνο cluster, είναι όπως είχαμε και σε άλλο σημείο αναφέρει, ότι κατά την διάρκεια λειτουργίας των εικονικών υπολογιστών, αυτοί επηρεάζονται από εξωγενείς παράγοντες (προγράμματα,updates, κλπ.) που τρέχουν στο background, με αποτέλεσμα την ομαλή λειτουργία του cluster.

6.2 **Προτάσεις περεταίρω μελέτης**

Έχοντας υπόψη μας τις δυνατότητες των clusters, μία καλή εφαρμογή θα ήταν σε Πανεπιστημιακά ιδρύματα και σχολές. Αν σκεφτούμε τις απαιτήσεις που έχουν τα σημερινά ιδρύματα σε ηλεκτρονικούς υπολογιστές και άλλου είδους περιφερειακά για servers, εργαστήρια, βιβλιοθήκες και άλλες χρήσης, χρειάζονται μερικές εκατοντάδες υπολογιστές το καθένα, με μία προοπτική χρήσεις για τουλάχιστον μια επταετία αν όχι περισσότερο. Κάτι τέτοιο όμως δημιουργεί σοβαρά προβλήματα, διότι η τεχνολογία εξελίσσεται με ιλιγγιώδεις ρυθμούς, σε επίπεδο hardware και software και οι απαιτήσεις των εφαρμογών σε εργαστηριακό επίπεδο κυρίως είναι αυξημένες, με αποτέλεσμα ένας μεσαίων δυνατοτήτων υπολογιστής να καταστείτε γρήγορα παρωχημένος μέσα σε τέσσερα χρόνια. Από την άλλη το κόστος αντικατάσταση εκατοντάδων υπολογιστών σε λιγότερο από πέντε χρόνια είναι κάτι το απαγορευτικό. Σε αυτό θα πρέπει να σκεφτούμε το κόστος των servers ενός ιδρύματος που υπολογίζεται σε αρκετές χιλιάδες ευρώ, ακόμα και αν η αντικατάσταση τους γίνεται ποιο αραιά. Τέλος πρέπει να υπολογίσουμε και τους εκατοντάδες αυτούς υπολογιστές που μετά την αντικατάστασή τους από τα ιδρύματα καταλήγουν στην καλύτερη των περιπτώσεων για κάποια δευτερεύων χρήση ή σε κάποιο κέντρο ανακύκλωσης, με λίγα λόγια εξοπλισμός πολλών χιλιάδων ευρώ παροπλίζεται μέσα σε λίγα χρόνια.

Μια καλή λύση σε αυτά τα προβλήματα θα μας έδινε η κατασκευή ενός cluster με όλους τους παρωχημένους τεχνολογικά υπολογιστές του ιδρύματος. Κάτι τέτοιο θα μας έδινε μεγάλη συνδυαστική επεξεργαστική ισχύ και αποθηκευτικό χώρο που σε πολλές περιπτώσεις θα μπορούσε να ξεπεράσει servers αρκετών χιλιάδων ευρώ, με ένα πάρα πολύ χαμηλό κόστος για την κατασκευή του. Το cluster με τους «παλιούς» υπολογιστές θα μπορούσε να είναι είτε cluster Υψηλής Διαθεσιμότητας (High Availability Cluster), είται Εξισσορόπησης Φορτίου (Load Balancing Cluster) και να χρησιμεύσει ως server του ιδρύματος με πολλές χρήσεις, ανάλογα με τις απαιτήσεις. Επίσης η αναβάθμιση και συντήρηση ενός τέτοιου cluster θα ήταν εύκολη και με μηδαμινό κόστος, καθώς ένα από τα μεγάλα πλεονεκτήματα των cluster είναι η επεκτασιμότητα, κάθε υπολογιστής που θα θεωρούταν παρωχημένος θα μπορούσε εύκολα να προστεθεί με τους υπόλοιπους υπολογιστές του cluster, αυξάνοντας τις δυνατότητές του.

Άρα στην ουσία έχουμε μία «ανακύκλωση» των διαθέσιμων πόρων, κάνοντάς τους ποιο παραγωγικούς με μηδαμινό κόστος και φιλικότερους προς το περιβάλλον.