Wednesday, December 27. 2006

Cluster Filesystem in SunCluster

Diese Erläuterung von Tatjana Heuser ist zu gut, um sie in einem Kommentar verdorren zu lassen:
Das Cluster Filesystem war im Prinzip die innovativste Idee am SunCluster 3.
Die Intra-Cluster Kommunikation laeuft ueber einen Object Request Broker. Dieser laeuft - zumindest bis 3.1 - mit eingeschalteten Debug Flags. (Fuer die gerade freigegebene 3.2 Release noch nachsehen). Wenn man das abschaltet verliert man erweiterte Debug-Moeglichkeiten, gewinnt jedoch Durchsatz. Auch mit Traffic Shaping/Einstellungen fuer das private Network laesst auch noch eine ganze Menge herausholen, auch dieses nicht ohne Preis/Nebeneffekte. Das Cluster Filesystem ist ein Corba basierter Dienst, der eben auch ueber diese Intra-Cluster Kommunikation laeuft. Daher ist der Durchsatz im Clusterfilesystem leider auch durch die
Qualitaet/Geschwindigkeit des Sun Cluster private Networks bedingt. Wer also weiterhin nur 2x 100 MBit Interfaces im private Network faehrt, limitiert damit die Intra-Cluster Kommunikations Performance und implizit damit auch das Cluster Filesystem.

HAStorage+ stellt im Prinzip eine Reinkarnation des einfachen Filesystem failover Mechanismus von Cluster 2.x dar. Dieser Mechanismus wird aus Performance Gruenden auch gerne im Sun Cluster 3.x verwendet. Wer also die Flexibilitaet des ClusterFS nicht braucht kann an der Stelle auf HAStorage+ zurueckgreifen. Wer hingegen die volle Flexibilitaet von Sun Cluster 3.x nutzen moechte, muss ins Backbone investieren.

Ansonsten ist die Node, die den physikalischen I/O macht, auch die Node, die den Masterservice faehrt. Die Servernode hat den physikalischen Handle, die Proxynodes greifen ueber das private Network (ORB) auf das Proxy Filesystem zu.Auf das im Kommentar auch noch erwähnte Buch zum Thema Sun Cluster bin ich wirklich schon gespannt.

Posted by Joerg Moellenkamp in German, Oracle at 10:30

Danke Tatjana und Jörg. Ich bin leider noch nicht weitergekommen. Gibt es irgend eine man page die das naeher beschreibt?

Wie heisst der Corba Daemon der die File Access Zugriffe annimmt. Wie heisst die client lib die die Umleitung macht. Wo kann ich quotas konfigurieren. Gibts ein status befehl auf dem ich sehen kann wer fuer einen gegebeneen globalen mount point der master ist, etc?

In der SC Doku steht dazu leider garnichts, ausser der -g option bei mount, und dass man ein DID device nehmen soll.

Gruss
Bernd

Gruss
Bernd
Anonymous on Jan 2 2007, 20:55

Die Frage ist nicht trivial zu benatworten.
Das Clusterfilesystem ist letztendlich ein Kernelmodul.
Die Node, die den physikalischen Mount hat, registriert im Orb den Clusterfilestem-Serverdienst. Alle Clusternodes, die auf dieses Filesystem zugreifen wollen, registrieren sich ueber einen IDL stub als Filesystem Proxy und fragen im Request Broker den vorher registrierten Server an.

Wenn nun dann deine Applikation so I/O empfindlich ist, gibt es im Primzip zwei Loesungen. Die eine heisst auf das Cluster Filesystem zu verzichten, und HAStorage+ zu verwenden. Wenn jedoch auf der anderen Seite das Cluster Filesystem genutzt werden soll, muss, wie Tatjana schon geschrieben hat, in mehr Leistung im Backbone investiert werden.
Es sind bis zu 16 Interfaces, 8 je Netz bei zwei Netzen im private Network moeglich. Das Cluster fuehrt eine entsprechende Lastbalance ueber alle Interfaces durch. An dieser Stelle kann man dann sinnvollerweise auch Gigabit Interfaces einsetzen, was den Durchsatz deutlich erhoeht.

Bitte denk daran, du musst die junctions, switches und switch types im private net neu definieren, falls die alte Konfiguration eine back-to-back Konfiguration war, aber das laesst sich alles im laufenden Betrieb umstellen. Wennn du mitlesen willst was im ORB passiert - dazu gibt es im Developer Package (Standard Bestandteil des Clusters) das Kommando orbadm. Mach Dich auf gute Unterhaltung gefasst, das Ding ist geschwaetzig .

Sorry, aber da nicht direkt ein Daemon fuer das Clusterfilesystem zustaendig ist, und ein traffic-shaping im Backbone deutliche Nebeneffekte fuer den regulaeren Clusterbetrieb bedeutet, ist eine einfache Antwort nicht moeglich. Aber wie Tatjana gesagt hat, Du kannst Die Debug Flags abschalten, das macht dann allerdings doch leider einen Reboot der Clusternodes notwendig, da Eintrag in die /etc/system. Um zu sagen ob das auch ueber einen einfachen Eingriff mit dem Kernel Debugger am lebenden System geht, bin ich nicht kompetent genug.
Anonymous on Jan 5 2007, 05:07

Danke Rolf, das hat mir schon mal weitergeholfen.

Ich habe aktuell kein konkretes Problem, ich wollte nur ein paar Messungen machen, um zu erkunden ob das global mount einen spührbar (negativen) Effekt hat.

Dazu wollte ich halt verstehen wie es funktioniert, und vor allem halt auch Diagnosemöglichkeiten zu haben, zum Beispiel "welcher der Knoten macht die physikalische IO".

Ich schau mich mal im ORB um.

Wenn ich richtig gelesen habe so setzt das HAStorage+ das wir einsetzen die Affinity der Device Gruppe auf die aktive Node.

Betrifft dies somit auch den Master Node der die IO für das globale Filesystem macht? Dann wären wir zumindest was den I/O Pfad angeht auf der sicheren Seite, und ich müsste nur noch die Delays durch den Cache/Orb betrachten (und die können nicht so gross sein, im Betrieb ist es ja ein read only filesystem).

Gruss
Bernd
Anonymous on Jan 5 2007, 06:00

BTW: rolf kannst du mir mal ne Mail schreiben, auf deinen Web Seiten habe ich keine Kontaktmöglichkeit gefunden (und der Admin-C deiner Domain bist du auch nicht).

Gruss
Bernd
Anonymous on Jan 5 2007, 06:32

Seufz,
ich warte immer noch auf ein vernuenftiges Buch zum Thema Sun Cluster. Besonders wenn man sich auf die Zertifizierung vorbereiten will. Docs.sun.com ist ja wunderbar. Aber die Inhalte werden ungluecklicherweise so trocken vermittelt, das einem das Lesen der umfangreichen Literatur manchmal doch extrem schwer faellt. Ohne Signalstifte, Kaffe und Zigaretten kommt man da leider nicht weit.
Anonymous on Oct 9 2008, 12:29

Aber mit mdb, DTrace, apptrace find, grep, ls und vi kommt man ueberraschend weit.....
Kommt demnaechst... Letztes Finetuning, Tonnenweise Tests, alles um Version 3.2 erweitert, zwar immernoch auf historischer Hardware aber mit neuem Solaris und diesmal mit Lektoriat:)
Ist aber nicht fuer Zertifizierungen geschreiben, eigentlich "nur" fuers Verstaendnis, also wenn man dann Samstag Abend schnell noch einen kleinen Update einfahren wollte und mit dem Ergebnis dessen was man da gebaut hat nicht wirklich "zufrieden" ist:(

Gruesse, Rolf
Anonymous on Apr 1 2009, 19:27