

6.6.4 Cluster Interconnect im Private Network

Zwischen den Knoten des RAC werden viele, meist sehr kleine Pakete über den Cluster Interconnect ausgetauscht. Kurze Latenzzeiten beim Austausch der Daten über den Cluster Interconnect sind daher für die Performance und ganz besonders für den Skalierungsfaktor im Cluster maßgeblich.

Generell können Ethernetkarten für den Cluster Interconnect genutzt werden. Um eine ausreichende Bandbreite zu erreichen, sollte zumindest Gigabit-Ethernet eingesetzt werden. Einige Hersteller bieten aber auch eigene, oftmals proprietäre Produkte an:

Interconnect	Übertragungsleistung (Gbit/s)	Latenzzeit (msec)
Fast-Ethernet	0,1	~ 3000
Gigabit-Ethernet	1	~ 3000
SCI (Scalable Coherent Interconnect)	10,66	~ 1500
Myrinet	2	~ 6,7
Memory Channel	1	~ 3
Hyper Fabric	4	~ 22
InfiniBand	2,5	< 500

Tabelle 6.4: Vergleichswerte verschiedener Interconnect-Technologien

Viele dieser Interconnect-Produkte erreichen den Latency-Level eines SMP-Busses¹. Technologien wie Memory Channel, SCI und Myrinet unterstützen Virtual Shared Memory im Sinne eines »Internode Memory-Adress-Mapping«. Aufgrund der Memory-Map-Architektur, die Zugriffsadressen knotenübergreifend in Adressbereiche des lokalen Arbeitsspeichers übersetzt, ist der Overhead bei der Internode-Kommunikation ähnlich dem beim Zugriff auf lokalen Speicher – und damit wesentlich geringer als bei Protokollen wie UDP und TCP.

Fast-Ethernet und Gigabit-Ethernet

In älteren RAC-Konfigurationen wird häufig noch Fast-Ethernet bzw. Ethernet eingesetzt. Da der Cluster Interconnect bei konkurrierenden Zugriffen auf Datenblöcke potenziell einen Flaschenhals bildet, sollte besser Gigabit-Ethernet verwendet werden. Dies wird auf allen von Oracle supporteten Plattformen für den Cluster Interconnect unterstützt.

Für einen Zwei-Knoten-Cluster kann die Verbindung über ein Crossover-Kabel hergestellt werden. Jedoch wird in der Regel ein passender Ethernet-Switch eingesetzt. Sowohl Netzwerkkarten als auch Switches sollten natürlich redundant ausgelegt sein.

¹ SMP-Bus: Prozessorbus in symmetrischen Multiprozessorsystemen

Memory Channel

Memory Channel ist ein High-Speed-Network Interconnect, der Applikationen in einem Cluster einen clusterweiten Adressbereich zur Verfügung stellt. Anwendungen im Cluster bilden diesen Adressbereich in ihren jeweils eigenen virtuellen Adressräumen als 8-Kbyte-Pages ab. Innerhalb der virtuellen Adressräume kann in der gleichen Weise wie bei Nutzung des lokalen Arbeitsspeichers gelesen und geschrieben werden. Alphabasierte HP-Cluster (ehemals Compaq) können diese Technologie nutzen.

Myrinet

Myrinet ist eine preisgünstige Paket- und Switch-Kommunikation, die hohen Performance Ansprüchen genügt. Myrinet unterstützt gebräuchliche Rechner- und Betriebssysteme und wird häufig in Linux-Clustern eingesetzt. Die Software wird übrigens als Open Source bereitgestellt. Myrinet nutzt Host Interfaces, die ein Steuerprogramm unter Umgehung des Betriebssystems (OS-Bypass) ausführen, und ermöglicht so Low Latency-Kommunikation über direkte Network Send-, Receive und Packet Puffer-Mechanismen.

Scalable Coherent Interconnect (SCI)

Scalable Coherent Interconnect (SCI) ist ein Sun High Performance Interconnect, der hohe Datenübertragungsraten und Low Latency Kommunikation bereitstellt. Anwendungen, die wie RAC einen hohen Bedarf an Cluster-Kommunikation über den Interconnect haben, skalieren mit SCI weit besser, als dies mit Ethernet der Fall ist. SUN SCI implementiert *Remote Shared Memory (RSM)* und umgeht damit den durch TCP/IP-Kommunikation verursachten Overhead. SCI ist u.a. für Sun Fire 4800- und 6800-Systeme verfügbar.

LLT und GAB der Veritas DBE/AC

Die Kommunikation der *Veritas Database Edition/Advanced Cluster (DBE/AC)* nutzt ein *Low Latency Transport-Protokoll (LLT)* sowie *Group Membership/Atomic Broadcast Services (GAB)*. LLT stellt eine Kernel-to-Kernel-Kommunikation bereit.

HP Hyper Fabric HMP

HP Hyper Fabric unterstützt Standard TCP/UDP über IP sowie HPs proprietäres Hyper Message-Protokoll (HMP). Hyper Fabric erweitert den TCP/UDP-Stack um Load Balancing über mehrere Netzwerkkarten. HMP in Verbindung mit OS-Bypass ermöglicht zudem Low Latency-Kommunikation sowie eine geringe CPU-Belastung im Rahmen des Datentransfers.

VIA

Virtual Interface Architecture (VIA) ist eines der verbreitetsten Low Latency-Protokolle auf Intel-Hardware. Dieses Memory-Map-Protokoll erlaubt Blockshipping über den Cluster Interconnect, ohne dass betriebssystemseitige Kerneloperationen notwendig sind. Zurzeit ist VIA nur mit proprietärer Hardware realisierbar. Jedoch sind spezielle Treiber, die VIA over Gigabit-Ethernet erlauben, in der Entwicklung.

VIA wurde designed, um die Performance verteilter Anwendungen durch Reduktion der Latenzzeiten zu optimieren. Dieses Ziel konnte erreicht werden, indem im Vergleich zu traditionellen Network Interface-Architekturen weniger System-Software in die Kommunikation involviert wurde.

Giganet

Giganet cLAN ist ein weiterer High-Speed, Low Latency Interconnect für Commodity Cluster. cLAN basiert auf VIA und erlaubt Bandbreiten bis 1,25 GB/s.

Infiniband

Infiniband kann sowohl für den Cluster Interconnect als auch für die Anbindung von Storage mittels Fibre Channel Gateway verwendet werden. Zudem ist Infiniband in der Lage, Shared Memory-Funktionalität bereitzustellen. Für den Einsatz von Infiniband wird ein HCA (Host Channel Adapter) benötigt, der mit einem Infiniband-Switch verbunden wird. Über Gateways sind Verbindungen zu IP- und Storage-Netzen möglich.

Infiniband wird im HPCC (High Performance Computing Clusters)-Umfeld eingesetzt, unterstützt mehrere Protokolle über einen HCA und ermöglicht Datenübertragungsraten von bis zu 3,75 GB/s Fullduplex. Der Einsatz dieser Technologie setzt ein PCIX-Bussystem (Erweiterung des PCI-Busses) voraus.

6.6.5 Public Network

Die Ansprüche an die Netzwerkkarten für das Public Network des Clusters unterscheiden sich nicht von jenen, die in einem Server eines Single Instanz-Systems eingesetzt werden. Typischerweise kommen Gigabit-Ethernetkarten zum Einsatz, die selbstverständlich ebenfalls redundant ausgelegt sein sollten.

6.6.6 Namen und Adressen des Private und Public Network

Einige Regeln sind bei der Konfiguration des Public und Private Network zu beachten. Zum einen müssen beide Interfaces unterschiedliche Namen tragen. Die Installationsroutine akzeptiert identische Namen für beide Interface-Typen nicht. Bei Nichtbeachtung kommt es sonst zu Fehlermeldungen während der Installation. Zudem sollten die IP-Adressen des Private Network in einem anderen Subnet als jene des Public Network liegen.

6.6.7 Voting Disk

Normalerweise nutzen die Clusterknoten den Cluster Interconnect, um zu kommunizieren. Bei konkurrierenden Zugriffen auf Datenblöcke wird über diesen auch das Sperrverhalten geregelt. Der Cluster Interconnect ist also enorm wichtig. Fällt er aus, so kann es bei konkurrierenden Zugriffen zu Blockkorruptionen kommen. Daher legt man in der Regel die private Netzwerkverbindung zwischen den Clusterknoten redundant aus. Fällt eine Netzwerkverbindung des Private Interconnect aus, kann über ein zweites Netzwerkinterface kommuniziert werden.

Fällt – aus welchen Gründen auch immer – das gesamte Private Network aus, so entsteht eine Situation, die als *Split Brain* bezeichnet wird: Ein Clusterknoten weiß nichts von der Existenz der restlichen Nodes und geht davon aus, der derzeit einzige Knoten im Cluster zu sein. Glauben nun mehrere Knoten, sie seien die derzeit einzigen im Cluster, kommt es zu unkoordinierten konkurrierenden Blockzugriffen und damit mit hoher Wahrscheinlichkeit zu Datenkorruptionen. Um diese Datenkorruptionen zu verhindern, nutzt der Oracle-Cluster ein spezielles Verfahren: Fällt das Private Network komplett aus, wird über eine Voting Disk das Eigentümerrecht am Cluster an genau einen Knoten vergeben. Dabei versucht jeder funktionsfähige Knoten, die Voting Disk zu reservieren. Der erste Knoten, dem dies gelingt, bleibt verfügbar. Alle anderen Knoten schalten sich automatisiert ab.

Die Voting Disk befindet sich ebenfalls auf dem Shared Disk-Subsystem. Die Voting-Informationen benötigen etwa 20 MB und können entweder auf Raw Device oder auf ein Cluster-Filesystem abgelegt werden. Besteht kein Zugriff auf die Voting Disk, so kann keiner der Knoten die Datenbank mounten.

6.7 Software-Architektur

6.7.1 Cluster Manager

In früheren Oracle-Versionen war es nötig, Cluster-Software eines Drittherstellers zu installieren, damit RAC überhaupt lauffähig war. Die Oracle RAC-Option nutzte dann die jeweilige Cluster-API des Drittherstellers für Dienstleistungen. Diese Software wird als *Cluster Manager* bezeichnet.

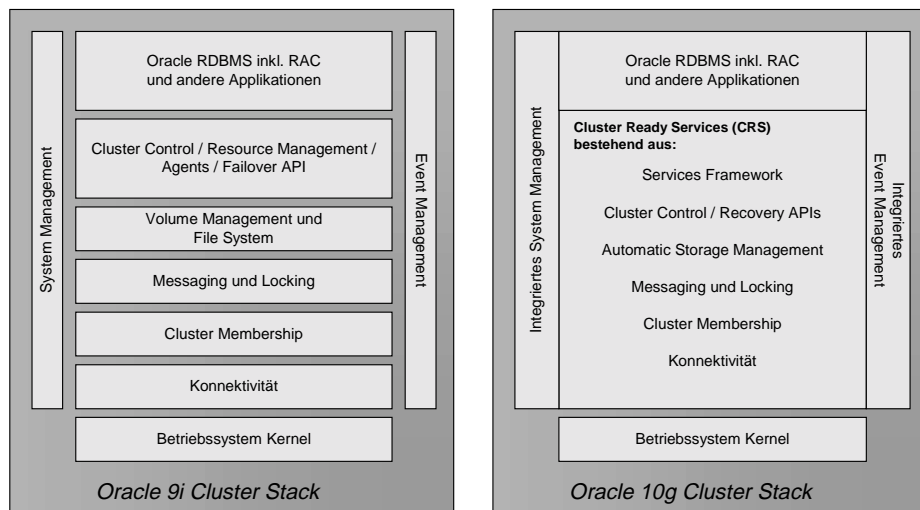


Abbildung 6.2: Cluster Stack in Oracle 9i im Vergleich zu Oracle 10g

Mit Oracle 9i wurde eine integrierte Cluster Management-Software für Linux und Windows eingeführt. Seit Oracle 10g ist diese nun auf allen von Oracle unterstützten Plattformen verfügbar. Gegenüber der Vorgängerversion wurde zudem die Anzahl der unterstützten Knoten auf 64 erhöht.

Der Oracle Cluster Manager liefert alle für den RAC notwendigen Clusterfunktionen. Hierzu zählt die Verwaltung der Node Membership sowie Group Services und Global Resource Management. Er kann auf allen Plattformen installiert werden, auf denen auch RAC lauffähig ist. Zusätzlich zum Oracle Cluster Manager kann auch weiterhin eine Cluster Management-Software eines Drittherstellers installiert werden. Notwendig ist dies jedoch nicht.

Bestandteil des Cluster Managers sind im Wesentlichen drei Services. Über diese wird die Kommunikation im Cluster gehandhabt. Es handelt sich dabei um

- ▶ die Cluster Synchronisation Services (CSS)
- ▶ die Cluster Ready Services (CRS) sowie
- ▶ den Event Manager

Auf Windows-Plattformen werden diese Funktionen über drei Services bereitgestellt, die diese Funktionen übernehmen: *OracleCSService*, *OracleCRService* sowie *OracleEVMService*. Diese drei Services ersetzen den Cluster Service *OracleCMService9i*, der unter Oracle 9i zum Einsatz kam. Auf Unix-Systemen finden sich drei Daemons anstelle der genannten Services: *ocssd*, *crsd* und *evmd*.

Alle drei Daemons bzw. Services generieren Tracefiles. Pfad und Name derselben sind Tabelle 6.5 zu entnehmen.

Dienst	Daemon	Tracefile
Cluster Synchronisation Services	CSS	<CRS Home>/css/log/ocssd<number>.log bzw. <CRS Home>/css/init/<node_name>.log
Cluster Ready Services	CRS	<CRS Home>/crs/init bzw. <CRS Home>/crs/<node name>.log
Event Manager	EVM	<CRS Home>/evm/log/evmdaemon.log sowie <CRS Home>/evm/init/<node_name>.log

Tabelle 6.5: Cluster Services und Tracefiles

6.7.2 Prozessarchitektur

Spezifische Prozesse des Oracle Real Application Cluster sind der Diagnoseability Daemon, der Lock-Prozess, der Lock Manager Daemon, der Lock Monitor sowie der Lock Manager Process. Während diese auf Windows-Plattformen in einer Thread-Architektur implementiert und nach außen hin nicht erkennbar sind, ist jeder dieser Prozesse auf Unix-Systemen in der Prozessliste sichtbar.