IT-Kompendium.

1. Rechnerarchitekturen
2. Betriebssysteme
3. Cluster
4. Speicherbusse
5. Festplatten und RAID
6. Speichernetzwerke
7. Magnetbandspeicher
8. Optische Speicher
9. Arbeitsspeicher
10. Kommunikation
11. Standards und Normen
12. Das OSI-Referenzmodell
13. Übertragungsverfahren und -techniken
14. Personal Area Networks – PANs
15. Local Area Networks – LANs
16. Metropolitan Area Networks – MANs
17. Wide Area Networks – WANs
18. LAN-Core-Lösungen
19. Eingabegeräte
20. Datenkommunikation
21. Terminals
22. Ausgabegeräte
23. Multimedia
24. Unterbrechungsfreie Stromversorgungen

im IT-Kompendium auf Seite 144 – 161

online auf unserer Website unter:
www.transtec.de, www.transtec.at, www.transtec.ch

1. Rechnerarchitekturen

1.1.9 Multi-Core-Prozessortechnologien

Das Jahr 2006 war für die Entwicklung von CPUs sehr bedeutsam. Durch die Einführung von Prozessoren mit mehreren Cores in einem Gehäuse wurde ein Leistungssprung möglich, wie es ihn schon lange Zeit nicht mehr gab. Durch diese Entwicklung wurde der Weg zum Parallel Computing freigemacht.

1.1.9.1 Begriffsklärungen

Single-Core-CPU (Central Processing Unit)

Bis zur Einführung von Dual-Core-CPUs von AMD und Intel® im Jahr 2004 konnte eine CPU durch die folgenden wesentlichen Funktionseinheiten beschrieben werden:

- ALU (Arithmetic and Logic Unit): dies ist das eigentliche Rechenwerk zur Abarbeitung der Maschinenbefehle.
- FPU (Floating Point Unit): der mathematische Koprozessor zur Bearbeitung von Fließkommazahlen.
- L1-Cache (Level 1 Cache): ein schneller Zwischenspeicher basierend auf SRAM, zur Zwischenspeicherung von Daten und Befehlen. Der L1-Cache läuft im optimalen Fall mit derselben Taktfrequenz wie die CPU.
- L2-Cache (Level 2 Cache): dieser Zwischenspeicher beschleunigt den Durchsatz der CPU durch vorausschauende Ladevorgänge aus dem relativ langsamen Arbeitsspeicher.

Hinzu kommen noch weitere Funktionseinheiten, wie Speicherinterface, Befehlsdecoder etc. Diese sollen hier nicht im Detail betrachtet werden.

Core

Ein Core besitzt im Prinzip denselben Aufbau wie eine vollständige CPU. Das Speicherinterface ist nicht Bestandteil eines Cores. Alle Cores müssen sich ein Speicherinterface zur Kommunikation mit dem Arbeitsspeicher teilen. Dies gilt auch für Kommunikationsinterfaces wie zum Beispiel den HyperTransport-Bus.

Bei der Integration des L2-Cache gibt es unterschiedliche Ansätze. Bei den Dual-Core Intel® Core™2 Duo Prozessoren (Conroe, Merom, Woodcrest) teilen sich beide Cores den L2-Cache. Beim AMD Opteron™ Dual-Core und Athlon™ 64 X2 Prozessor dagegen besitzt jeder Core seinen eigenen L2-Cache.

1.1.9.2 Warum mehrere Cores in einer CPU?

Bis zur Vorstellung der ersten Opteron™ Dual-Core-CPUs von AMD im Jahr 2005 wurde die CPU-Leistung durch Verbesserung der Architektur und Erhöhung der Taktfrequenz gesteigert.

Durch die stetige Verbesserung der Produktionsprozesse und damit verbundene Verkleinerung der Strukturen auf der CPU war die Erhöhung der Taktfrequenz relativ leicht. Im Jahr 2001 lagen die Taktfrequenzen typischer CPUs bei 1 GHz. Beispiele hierfür sind der Intel® Pentium® 4 Willamette Core und der AMD Athlon™ Thunderbird, die in 180 nm Fertigungstechnik hergestellt wurden.

Bis zum Jahr 2005 konnte die Taktfrequenz durch Fertigungsprozesse mit verkleinerten Strukturen (90 nm und 65 nm) bis auf ca. 3,5 GHz gesteigert werden.

Die optimistischen Voraussagen, dass sich die Taktfrequenz von CPUs bis zu 5 GHz und mehr steigern lassen könnte, wurden allerdings nicht erfüllt. Der Grund hierfür ist die Abhängigkeit der Leistungsaufnahme P von der Spannung U und der Taktfrequenz f. Es gilt:

$$P = C_{CPU} U^2 f$$

Diese Beziehung gilt allgemein für CMOS-Schaltungen, wie sie heute die Grundlage für jeden Prozessor sind. Hierbei ist C_{CPU} eine Proportionalitätskonstante die von der CPU-Architektur und dem verwendeten Fertigungsprozess abhängt. Durch eine Verkleinerung der Strukturbreiten bei der Fertigung kann der Faktor C_{CPU} verringert werden. Hier sind durch den Übergang von 65 nm auf 45 nm und darunter auch in der Zukunft noch weitere Verbesserungen möglich.

Die typische Leistungsaufnahme einer CPU mit 1 GHz Taktfrequenz lag bei 60 W. Die hochgetakteten CPUs des Jahres 2005 erreichen teilweise über 100 W. Die damit verbundene thermische Belastung der CPU und des Gesamtsystems ist nur noch sehr schwer über eine Luftkühlung zu bewältigen. CPU-Temperaturen von 70 °C sind keine Seltenheit.

Hierunter leidet auch die Lebensdauer der CPU, da sich Effekte wie die Elektromigration durch diese hohen Temperaturen verstärken (Arrhenius-Gesetz).

Da eine Steigerung der Taktfrequenz auch eine Erhöhung der Spannung U bedingt, führt dies wegen des Faktors U^2 zu nicht mehr zu beherrschenden Leistungsaufnahmen P. Streng genommen muss also die Spannung U als Funktion von f betrachtet werden.

Der Ausweg aus diesem Dilemma ist die Einführung von zwei oder mehr Cores pro CPU, die im Vergleich zu Single-Core-CPUs mit verringerter Taktfrequenz laufen. Dadurch wird es möglich, die Leistungsaufnahme wieder auf erträgliche Werte zu reduzieren.

1.1.9.3 Technische Ausführungen von Multi-Core-CPUs

Moderne Fertigungsprozesse mit 65 nm Strukturbreite machen es heute möglich, auch in CPUs für Mobilgeräte bereits Dual-Core-Technik einzusetzen. Die High End Desktop und Server CPUs von Intel® verfügen bereits über vier Cores.

Diese Quad-Core-CPUs bieten einen Leistungssprung gegenüber den besten Single-Core-CPUs, wie er seit einigen Jahren nicht mehr zu beobachten war.

Die momentan verwendeten 65 nm und 90 nm Prozesse erlauben es, zwei Cores pro Chip zu fertigen. Die Intel® Quad-Core-CPUs Core™2 Extreme QX6700 und Xeon® 5300 besitzen zwei Dual-Core-Chips in einem LGA-Gehäuse. Diese Technik ist in der Fertigung aufwändig und entsprechend kostenintensiv, wird aber vom Prozessorhersteller verfolgt, um möglichst früh ein Quad-Core-Design an den Markt zu bringen.

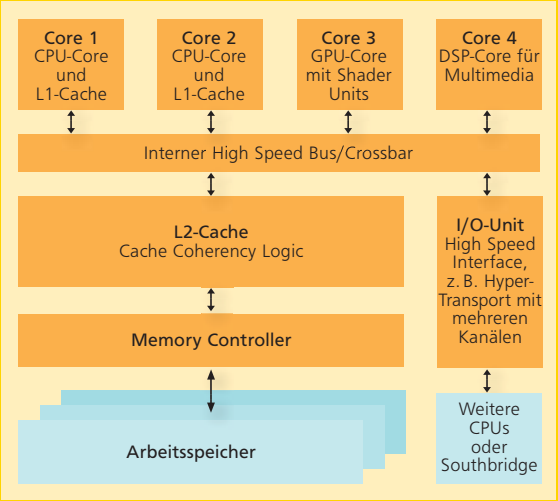
Langfristig kann ein Multi-Core-Design nur aus einem Chip pro Gehäuse bestehen, da sich nur so die Kosten minimieren lassen. Damit werden aufwändige und fehleranfällige Bondings reduziert.

Bei der Architektur des L2-Cache gibt es unterschiedliche Ansätze. Intel® verwendet einen shared L2-Cache, der zwei Cores gemeinsam zur Verfügung steht. Dies ist bei Programmen, die nur aus einem Thread bestehen, von Vorteil, da diese einen großen Cache benutzen können, der dann einem Core komplett zur Verfügung steht. AMD stellt jedem Core einen kleineren L2-Cache zur Verfügung. Dafür entfällt hier die schwierige Verwaltung des L2-Cache für beide Cores. Die Verwaltungsprobleme eines gemeinsam verwendeten Caches werden sich bei zukünftigen Multi-Core-CPUs noch vergrößern.

Die Kommunikation von zwei Cores findet bei Intel® über den Frontside Bus statt. Bei den AMD CPUs sind die Cores über einen Crossbar Switch verbunden. Dieser Ansatz wird sich langfristig durchsetzen, da die Cores damit nicht das externe Bussystem belasten. Zudem kann der Crossbar Switch auch mit höheren Taktraten betrieben werden als das externe Bussystem.

Die kommende AMD Quad-Core-CPU „Barcelona" vereint alle vier Cores auf einem Chip. Die Kommunikation übernimmt wieder ein Crossbar Switch. Als Novum ist hier der gemeinsam verwendete L3-Cache zu sehen.

Es sei noch angemerkt, dass die Multi-Core-Technik keine Erfindung von AMD oder Intel® ist. Am Massachusetts Institute of Technology wurde bereits 1997 ein Chip mit 3 Cores entwickelt.



Eine zukünftige Multi-Core-CPU mit zwei spezialisierten Cores für Grafik und Media-Streams, beispielsweise für HDTV-Decoding/-Encoding. Der L2-Cache wird in diesem Fall von allen Cores gemeinsam genutzt.

1.1.9.4 Softwaretechnische Besonderheiten von Multi-Core-CPUs

Eine Single-Core-CPU basiert auf der Von-Neumann-Architektur. Dies gilt deshalb auch für einen Core einer Multi-Core-CPU. Eine Von-Neumann-Maschine kann Befehle nur sequentiell abarbeiten. Die Leistung kann in erster Linie erhöht werden durch breitere Register (zum Beispiel 64 Bit-Register statt 32 Bit) und durch die Erhöhung der Taktfrequenz.

Zudem gibt es noch die Möglichkeit, durch Verbesserungen der Architektur die Anzahl von Taktzyklen pro Befehl zu verringern. Ein Beispiel hierfür ist die RISC-Technik, bei der nur wenige Maschinenbefehle existieren, die aber sehr schnell abgearbeitet werden können. Eine andere Technik sind lange Befehls-Pipelines. Damit werden möglichst viele CPU-Komponenten gleichzeitig beschäftigt und Leerlaufzeiten minimiert.

Techniken wie VLIW (Very Long Instruction Word) befinden sich auch nach Jahren der Entwicklung noch in einem nicht voll praxistauglichen Stadium.

Wie bereits eingangs beschrieben, sind allerdings die technischen Möglichkeiten zur weiteren Leistungssteigerung durch diese Maßnahmen begrenzt.

Die Multi-Core-Technik setzt mehrere Cores mit Von-Neumann-Architektur zu einer CPU zusammen, die im Prinzip eine Leistungssteigerung proportional zu der Anzahl der Cores bietet. In der Praxis gibt es allerdings einige Umstände, die eine solch ideale Skalierung verhindern. Hier die zwei wichtigsten Faktoren:

Bandbreite des Speicherinterface

Eine Multi-Core-CPU kann nicht jedem Core ein eigenes Speicherinterface zur Verfügung stellen. Bei speicherintensiven Anwendungen ist das Speicherinterface der begrenzende Faktor. Beispiele für diese Anwendungen sind Computersimulationen im Bereich der CAE, Fluid- und Aerodynamik. Dort werden Modelle im Bereich von einigen GByte Größe berechnet.

Hier kann ein System mit zwei Dual-Core-CPUs gegenüber einem System mit einer Quad-Core-CPU im Vorteil sein, da dort jeder CPU eine eigene Speicherbank zur Verfügung steht.

Ein größerer L2-Cache bringt hier Vorteile, allerdings belegt dieser sehr viel Fläche auf dem Chip. Der L2-Cache wird immer mit SRAM (Static RAM) ausgeführt, bei dem je nach Auslegung 4 oder 6 Transistoren pro Bit eingesetzt werden. Damit sind dem Ausbau des L2-Cache entsprechende Grenzen gesetzt.

Multithreading

Ein Programm muss in mehrere parallel ausführbare Threads aufgetrennt werden können, damit es die zur Verfügung stehenden Cores auch auslastet. Ein Programm, welches zwei Threads besitzt, kann gut auf einer Dual-Core-CPU eingesetzt werden. Auf einer Quad-Core-CPU wird dagegen keine Leistungssteigerung möglich.

Programmierer sind heute bestrebt, die Multithreading-Fähigkeit unterschiedlichster Programme zu verbessern. Diese Aufgabe erweist sich allerdings als sehr schwierig, entsprechend langsam ist der Fortschritt auf dem Gebiet. Einige Algorithmen lassen es nur sehr schwer zu, Programme mit gutem Multithreading zu schreiben. Ein Beispiel ist 3D-CAD-Software, die auf der Methode der Constructive Solid Geometry basiert. Aus diesem Grund laufen viele dieser Programme nur als ein einziger Thread ab.

Dagegen skalieren Programme aus dem Bereich der photorealistischen Grafik sehr gut mit der Anzahl der Cores. Hier kommt der Ray Tracing Algorithmus zum Einsatz, mit dem die Ausbreitung von Licht in einer Szenerie simuliert wird. Da sich Lichtstrahlen ohne Wechselwirkung beliebig überlagern, können im Extremfall alle Pixel einer Szenerie gleichzeitig berechnet werden.

Ein Beispiel hierfür ist der Open Source Raytracer Povray. Der Kern des Programms hat nur einige hundert KByte und passt deshalb in den L2-Cache aller aktuellen Multi-Core-CPUs. Die Szeneriedaten basieren auf der Constructive Solid Geometry und sind daher sehr kompakt. Das Speicherinterface ist damit kein Engpass.

Die Leistung von Povray ist proportional zu der Anzahl der Cores und deren Taktfrequenz. Es handelt sich hier um den Idealfall einer Multithreading-Anwendung.

1.1.9.5 Besonderheiten des Betriebssystems

Der Scheduler des Betriebssystems soll laufende Prozesse so auf die einzelnen Cores verteilen, dass diese optimal ausgelastet werden. Ein Prozess besteht aus einer Instanz eines laufenden Programms, Speicherbereichen, Variablen, Sicherheitsattributen und dem Zustand der CPU-Register. Sind nun zahlreiche Prozesse zu verwalten, so kann der Scheduler oft nicht die optimale Verteilung ermöglichen. Da einzelne Prozesse multithreaded ausgelegt sein können, ergeben sich für den Scheduler noch weitere Probleme bei der effizienten Aufteilung auf die Cores.

So gab es beispielsweise unter Windows® XP Probleme bei der Nutzung der ersten Dual-Core-Prozessoren. Zukünftige Multi-Core-Prozessoren werden dem Betriebssystem bis zu 32 und mehr Cores zur Verfügung stellen. Eine derartige Konfiguration ist mit einem 8-Core Prozessor in einer 4-CPU-Konfiguration zu realisieren. Dies dürfte im Serverumfeld ab 2009 bereits machbar sein.

Damit kommen auf die Entwickler von Betriebssystemen auch entsprechend große Probleme bei der optimalen Nutzung dieser Cores zu. Es bleibt zu hoffen, dass die Softwareentwicklung hier mit der Hardware Schritt halten kann.

1.1.9.6 Weitere Multi-Core-Architekturen

Cell Broadband Engine

Eine wichtige Entwicklung im Bereich Multi-Core ist die Cell Broadband Engine (CBE), die aus einer Zusammenarbeit von IBM®, Toshiba® und Sony® hervorgegangen ist. Die CBE kommt in der Sony® PlayStation®3 zum Einsatz, aber es gibt bereits einige Zusatzkarten für PCs, die auch damit ausgestattet sind. Diese nutzen die sehr hohe Fließkommaleistung der CBE, wie sie beispielsweise für numerische Berechnungen im Computer Aided Engineering Bereich nötig ist.

Die CBE unterscheidet sich in vielen Punkten von den Multi-Core-Designs von Intel® und AMD, da es sich nicht um gleichberechtigte Cores handelt.

Die CBE besteht aus den folgenden Funktionseinheiten:

- PPE (Power Processor Element): eine auf dem PowerPC® basierende CPU
- 8 x SPE (Synergistic Processing Element): für Integer- und Floating Point Arithmetik optimierte Cores
- EIB (Element Interconnect Bus): ein schnelles Bussystem für die interne Kommunikation der SPEs und PPE
- Memory Controller, basierend auf Rambus XDRAM-Technik

Die PPE ist die Schnittstelle zum Betriebssystem, da es sich hier um eine komplette PowerPC® CPU handelt. Numerische Befehle werden an die acht SPEs weitergegeben, die zusammen bis zu 200 GFLOPS mit einfacher Genauigkeit und immerhin noch 14 GFLOPS bei doppelter Genauigkeit erreichen. Dies macht die CBE zu einem leistungsstarken FPU-Beschleuniger. So hat IBM® Pläne für Blades, basierend auf der CBE, die im Bereich des Supercomputing zum Einsatz kommen sollen.

Die SPEs ähneln den Vektorprozessoren, wie sie zum Beispiel in den Cray-Supercomputern eingesetzt wurden. Sie können ihre volle Leistung nur entfalten, wenn der Programmcode speziell an ihre Architektur angepasst wird.

Die nächste Generation der CBE mit 65 nm Technologie wird SPEs mit Optimierung für doppelte Genauigkeit besitzen; damit ergeben sich hier noch bessere Möglichkeiten, sehr leistungsstarke Beschleuniger zu entwickeln. Für die meisten Anwendungen aus der Numerik ist einfache Genauigkeit nicht ausreichend.

NVIDIA GeForce® 8800 GPU

Moderne Grafikkarten besitzen mehrere Shader Units, die als spezialisierte Floating Point Unit (FPU) Cores betrachtet werden können. Bei der GeForce® 8800 Graphics Processing Unit (GPU) von NVIDIA® gibt es jetzt keine feste Zuordnung dieser 128 Shader Units nach Vertex oder Pixel Shadern mehr.

Jede Shader Unit kann Berechnungen mit einfacher Genauigkeit durchführen. Es gibt Routinen in C, die typische Anwendungen aus der Numerik bereits sehr gut parallelisiert auf diese Shader Units abbilden können. Hierzu zählen Algorithmen aus der Bildbearbeitung wie die schnelle Fourier-Transformation (FFT). Bei den hochoptimierten BLAS3-Routinen für Matrixoperationen liegt die Rechenleistung der GeForce® 8800 GPU im Bereich von 100 GFLOPS.

Die Shader Units sind noch weit von der Flexibilität einer regulären CPU entfernt. Sie besitzen nur einen sehr eingeschränkten Befehlssatz, der die Programmentwicklung erschwert. Programme mit zahlreichen Verzweigungen und bedingten Sprüngen eignen sich nicht für die Abbildung auf die Shader Units. Wiederum ist die Beschränkung auf einfache Genauigkeit ein Nachteil, der aber in kommenden Generationen von GPUs mit doppelter Genauigkeit verschwinden wird.

1.1.9.7 Ausblick

Die Entwicklung von Multi-Core-CPUs kann zurecht als revolutionärer Schritt in der CPU-Entwicklung bezeichnet werden. Im Lauf des Jahres 2007 werden sich Dual-Core-Architekturen auch im Entry Level durchsetzen, während das Zeitalter der Quad-Core-Prozessoren bei Servern und Workstations anbricht.

Rechnerarchitekturen

Die Einführung der 65 nm Technologie wird 2007 umgesetzt worden sein. Der Umstieg auf die 45 nm Technologie wird vom Marktführer Intel® bereits für 2008 avisiert. Damit stehen den CPU-Entwicklern bei CPUs für Server bis zu einer Milliarde Transistoren pro Chip zur Verfügung. Eine Fab zur Fertigung von 65 nm CPUs kostet über 2 Milliarden Dollar. Nur die größten Halbleiterhersteller bzw. Kooperationen von Firmen können sich diese Anlagen leisten.

Hier die wichtigsten Konsequenzen dieser Entwicklung:

- Die weitere Verkleinerung der Transistoren erhöht die Leckströme erheblich. Diese sind heute bereits in derselben Größe wie die dynamische Leistung durch Schaltvorgänge.
- Heutige CPUs arbeiten mit Core-Spannungen von 1,1 V. Eine weitere Senkung ist nur schwer möglich und wird bei Werten von 0,7 V stagnieren. Gemäß $P = C_{CPU}U^2f$ ist der Verringerung der Leistungsaufnahme durch die Absenkung der Spannung damit eine enge Grenze gesetzt.

In der Vergangenheit konnte mit jedem Sprung auf kleinere Strukturbreiten auch die Taktfrequenz erhöht und die Leistungsaufnahme gesenkt werden. Aufgrund physikalischer Grenzen der Transistorentwicklung sind solche positiven Entwicklungen in Zukunft nicht mehr möglich. Dieses Problem wird allgemein als „Power Wall" bezeichnet.

Der Ausweg für CPU-Designer ist nur über den Ausbau der Multi-Core-Technologie möglich. Die Anzahl der Cores pro CPU dürfte bei Einführung der 32 nm Technologie auf 50 bis 100 ansteigen. Um die Leistungsaufnahme noch zu beherrschen, werden diese Cores keine wesentlich höheren Taktfrequenzen aufweisen können als heute bereits üblich. Die begrenzte Anzahl an Pins eines CPU-Gehäuses ermöglicht nur eine sehr eingeschränkte Verbesserung der Anbindung an den Hauptspeicher; somit werden zukünftige Multi-Core-CPUs durch diesen Faktor in der Leistung limitiert. Das klassische Cache-Konzept erweist sich hier ebenfalls als Engpass, da es nicht mehr ausreicht, um die Cores schnell genug mit Daten zu versorgen.

Für die Software- und Betriebssystem-Entwickler bricht hiermit zwangsläufig die Ära des Parallel Programming an, bei der eine Leistungssteigerung nur noch zu erzielen ist, wenn die Cores gleichzeitig ausgelastet werden können. Dies macht die Entwicklung von massiv multithreaded-programmierter Software notwendig. Diese Entwicklung ist wesentlich komplexer und teurer als singlethreaded Software, da sie die Entwicklung neuer Algorithmen notwendig macht. Der Fortschritt auf diesem Gebiet hinkt der CPU-Entwicklung hinterher.

Hier könnte das Stream Programming einen Ausweg darstellen. Dabei wird durch spezialisierte Compiler versucht, die Kommunikation zwischen einzelnen Funktionsblöcken auf dem Chip zu minimieren und alle benötigten Daten vor Beginn einer Operation bereits aus dem Hauptspeicher in lokale Zwischenspeicher zu laden. Dies vermeidet „Cache Misses", die zu einer Unterbrechung des Datenstroms zu einem Core führen. Bei GPUs kommt das Stream Programming bereits zum Einsatz.

Die durch Multi-Core-CPUs mögliche Leistungssteigerung wird damit entscheidend von der Verfügbarkeit der entsprechenden Software bestimmt. Für Anwendungen aus der numerischen Simulation gibt es bereits sehr gut parallelisierte Software. Der technisch wissenschaftliche Bereich wird daher von der Multi-Core-Technik zuerst profitieren.

Weiterhin ist es denkbar, dass die einzelnen Cores einer CPU unterschiedliche Architekturen aufweisen. Die „Fusion"-Technologie von AMD verfolgt diesen Ansatz. Dabei sollen auch GPU-Cores auf dem CPU-Chip zum Einsatz kommen, um die Grafikkarte zu entlasten oder zu ersetzen.

Hier einige Beispiele für zukünftige Cores:

- DSP-Cores (Digital Signal Processing) für die Bearbeitung von Audio- und Video-Streams in Echtzeit
- FPU-Cores (Floating Point Unit) für rechenintensive Aufgaben. Diese können die SPEs der CBE als Vorbild haben
- TCP/IP-Offload Engines zur Entlastung der CPU-Cores bei hoher Netzwerklast und bei iSCSI-Applikationen

Weiterhin können einige Cores auch nur als „Sea of Gates" ausgeführt werden, vergleichbar den heutigen FPGAs. Die einzelnen Gates können per Software zu einem Core zusammengefasst werden. Damit ergibt sich eine bisher noch nicht erreichte Flexibilität. Diese FPGAs erreichen im Moment nur bis 1 GHz Taktfrequenz, aber durch die Optimierung des Layouts der Gates an die Applikation kann trotzdem eine hohe Leistung erzielt werden.

1.1.10 Virtualisierung mit Intel® VT

Intel® VT steht für Intel Virtualization Technology (Codename Vanderpool) und bezeichnet im Prozessor implementierte Virtualisierungsfunktionen. Diese dienen dazu, Virtualisierungssoftware schlanker und robuster zu gestalten, indem wichtige Funktionen in die Hardware verlagert werden. Diese Techniken machen Servervirtualisierung schneller und flexibler, zum Beispiel setzen bestimmte softwarebasierende Lösungen eine Anpassung des virtualisierten Betriebssystems voraus. Mit der VT-Implementierung entfällt diese Beschränkung – so kann dann beispielsweise ein unmodifiziertes Windows® System als Gast auf einem Linux-Server laufen.

1.1.10.1 Servervirtualisierung

Warum Virtualisierung?

Virtualisierung, insbesondere die Servervirtualisierung, eröffnet Unternehmen eine Vielzahl großer Vorteile und setzt dabei eine enge Verzahnung der benötigten Hardware- und Softwarekomponenten für optimale Nutzung und Performance voraus.

Servervirtualisierung kann Unternehmen helfen, in großem Umfang Kosten durch deutlich verbesserte Ausnutzung der Hardware und durch Einsparungen bei Administration, Rechnerausstattung und Betrieb zu reduzieren.

Indem virtualisierte Betriebssysteme von der darunterliegenden Hardware unabhängig gemacht werden, verhilft Servervirtualisierung dem Betrieb von Servern und Anwendungen zu hoher Flexibilität und Sicherheit, denn ein Gastsystem präsentiert sich wie ein in sich geschlossenes Objekt: Es kann zwischen Servern verschoben werden, es kann kopiert und rasch wiederhergestellt und einfach gesichert werden. Unternehmen können so schneller auf sich ändernde Bedürfnisse und neue Situationen reagieren.

Wie funktioniert Servervirtualisierung?

Servervirtualisierungslösungen ermöglichen es, einen Rechner in voneinander isolierte Partitionen oder Container – die sog. virtuellen Maschinen – zu unterteilen. Dabei kommt meist eine Softwareschicht zum Einsatz, welche je nach Architektur direkt auf der Hardware oder wie eine Applikation auf dem Betriebssystem läuft und eine Ablaufumgebung für die virtuellen Maschinen bereitstellt.

Diese Schicht – meist als Virtual Machine Monitor (VMM) oder Hypervisor bezeichnet – ermöglicht es mehreren verschiedenen Betriebssystemen, sich die Hardware-Ressourcen eines Rechners zu teilen. Die Software sorgt dafür, dass die verfügbaren Kapazitäten von RAM, Prozessor, I/O, DMA und alle übrigen relevanten Komponenten transparent auf die Gastsysteme aufgeteilt werden.

Ein VMM hat primär folgende Aufgaben:

- Isolation: Er muss eine sichere Isolierung der virtuellen Maschinen voneinander innerhalb der geteilten physischen Ressourcen garantieren:
  - Datensicherheit und Vertraulichkeit sowie Konsistenz: Die VMs dürfen keinen gegenseitigen lesenden oder schreibenden Zugriff haben.
  - Stabilität: Eine amoklaufende virtuelle Maschine darf andere VMs nicht destabilisieren oder gar zum Absturz bringen.
- Effizienz: Die virtuelle Maschine darf aufgrund der Virtualisierung keinen unangemessenen Overhead produzieren, sondern sollte annähernd so schnell laufen wie auf der blanken Hardware („bare metal").

Die Anforderungen an virtuelle Maschinen stellen eine große Herausforderung für die Entwicklung leistungsfähiger Virtualisierungssoftware sowie der zugehörigen Verwaltungswerkzeuge dar. Vor allem die Performance ist dabei eine der zentralen Hürden.

Dies hängt insbesondere mit den Unzulänglichkeiten der x86-Prozessorplattform zusammen, welche nicht für Virtualisierung konzipiert war, sondern prinzipiell vom Paradigma „ein Rechner = ein Betriebssystem" ausgeht. Erst die neueste, seit 2006 verfügbare „virtualisierende" VT-Prozessorgeneration kann die Software in ihren Virtualisierungsvorhaben gezielt unterstützen.

Bisherige Virtualisierung ohne Hardwareunterstützung

Traditionell gehört die hardwarebasierende Virtualisierung in den Bereich der Mainframes und reicht bereits über 40 Jahre zurück. Aufgrund immer leistungsfähiger werdender kostengünstiger Server-Hardware und einer immer größeren Zahl mächtiger Softwarevirtualisierungslösungen wird Virtualisierung immer mehr zum Gemeingut, aber auch immer mehr zur Notwendigkeit für moderne IT-Umgebungen.

Intel® VT adressiert die Unzulänglichkeiten der alten Hardwarearchitekturen, denn die x86-Plattform hat ihre Wurzeln im Desktop-PC. Dieser muss traditionell nur vergleichsweise geringe Arbeitsleistungen erbringen und „hostet" prinzipiell nur ein Betriebssystem.

So ist es nicht verwunderlich, dass die x86-Prozessorarchitektur auch eine Reihe von Befehlen kennt, welche der softwarebasierenden Virtualisierung hinderlich sind.

Bislang mussten Virtualisierungslösungen daher komplett in Software realisiert werden. Dabei leidet entweder die Performance oder die Praktikabilität.

Wo liegen die Herausforderungen?

x86-Prozessoren arbeiten im Protected Mode mit einer Ring-Architektur (privilege levels), die es ermöglicht, dass der Betriebssystem-Kernel sowie Applikationen in unterschiedlichen, voneinander geschützten Bereichen ausgeführt werden: Kernel-Code zumeist im maximal privilegierten Ring 0, Benutzerprozesse im unprivilegierten Ring 3.

Um ein virtualisierendes System realisieren zu können, müssen zum Schutz des in Ring 0 laufenden Hypervisors die virtuellen Maschinen in einen anderen Ring verlagert werden. Ein typischer Hypervisor wie z. B. Xen™ sieht die in Abb. 1 dargestellte Architektur vor.

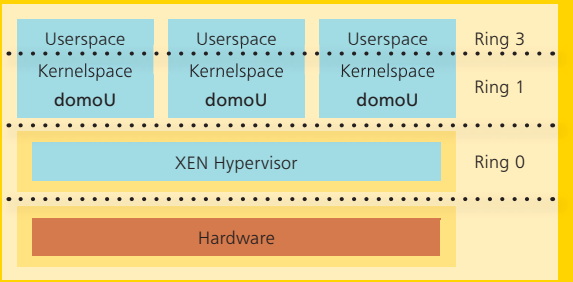


Abb. 1: Ring-Architektur

Die Gäste werden in den Ring 1 verschoben und können damit keine privilegierten Instruktionen mehr ausführen. Versuchen sie es (zwangsläufig) trotzdem, löst der Prozessor eine Exception (interner Fehlerstatus) aus, die dann vom Hypervisor behandelt werden kann, z. B. durch Beenden des Gastes oder Emulation des betreffenden Befehls.

Aus Sicht der Memory Management Unit (MMU), die nur die beiden Privilegstufen user und system kennt, laufen der Gast im Ring 1 und der Hypervisor im Ring 0 beide mit system-Privilegien. Damit könnte der Gast also Speicher beschreiben, der dem Hypervisor gehört. Um sich davor zu schützen, verwenden manche Hypervisoren zudem Segmente und setzen deren Limits so, dass der Gast nicht in den Bereich des Hypervisors hineinkommt.

Weil die Gäste im weniger privilegierten Ring 1 laufen, kann der Hypervisor nicht alle verfügbaren Instruktionen ausführen, ohne dass eine Prozessor-Exception die Folge ist, welche vom VMM zeit- und ressourcenaufwändig abgearbeitet werden muss.

Ingesamt gibt es in der x86-Architektur 17 solcher Instruktionen, mit denen der Hypervisor „kreativ" umgehen muss. Diese Instruktionen schließen auch solche ein, die der Gast zwar ausführen kann, weil die Instruktion an sich keine Ring 0-Privilegien braucht, die sich aber dann einfach anders verhalten, indem sie

- nicht den beabsichtigten Effekt erzielen (Beispiel: popf ignoriert das Interrupt Flag, wenn es nicht im Ring 0 läuft)
- andere Informationen zurückliefern als ihr privilegiertes Gegenstück
- Informationen liefern, die user nicht haben sollte

Der Hypervisor als kontrollierende Instanz muss nun das korrekte Verhalten als Reaktion auf die Instruktionen des Gastes erreichen. VMware® (Server, Workstation, GSX) beispielsweise filtert diese (zeit und ressourcen-)aufwändig mit bis zu 40 % Performanceverlust heraus, während der Paravirtualisierungsansatz von Xen™ das gewünschte Verhalten durch „Software-Interrupts" (sog. Hypercalls) erreicht und dabei deutlich performanter ist, aber eben eine Anpassung des Betriebssystemkernels des Gastes erforderlich macht.

1.1.10.2 Architekturen der Servervirtualisierung

Aufgrund der bislang mangelnden hardwareseitigen Unterstützung sind im Laufe der Zeit sehr verschiedene Modelle und unterschiedliche Verfahren entwickelt worden, um Servervirtualisierung zu realisieren. Die meisten Systeme basieren dabei auf einem „privilegierten" Betriebssystem, welches die volle Kontrolle über die Hardware-Ressourcen hat und diese auf die „Gäste" aufteilt.

Betrachtet man die gängigsten Software-Produkte, so lassen sich im Wesentlichen drei Kategorien unterscheiden:

- **„Vollständige Virtualisierung"** („Full Virtualization") durch Virtualisierung einer Hardwareumgebung. Der Virtual Machine Monitor läuft auf einem Betriebssystem.
- **Virtualisierung des Betriebssystems** mit einem einzigen Betriebssystem-Image bzw. einem Kernel (daher auch „Single Kernel Image" oder „SKI").
- **Paravirtualisierung:** Der Hypervisor/VMM läuft direkt auf der Hardware.

Vollständige Virtualisierung

Bei der vollständigen Virtualisierung (manchmal auch als „Partitionierung" bezeichnet) wird die Tatsache, dass eine Virtualisierung stattfindet, vollständig vor dem Gast verborgen. Der Virtual Machine Monitor (VMM) läuft dabei auf einem Host-Betriebssystem – meist im Userspace als Applikation – und fungiert wie ein Übersetzer zwischen Gast und Wirt.

Bekannteste Vertreter dieser Kategorie sind VMware® (Workstation, Server, GSX) sowie Microsoft® Virtual PC/Server.

Virtualisierung auf Betriebssystemebene

Dieses Konzept realisiert Virtualisierung, indem virtuell mehrere Instanzen ein und desselben Betriebssystems laufen. Hier wird also das Betriebssystem virtualisiert, nicht die Hardware. Das Host-Betriebssystem erzeugt weitere Instanzen seiner selbst, indem Benutzerprozesse und Applikationen in sog. „Ressourcen-Containern" gruppiert werden. Alle virtuellen Maschinen sind vom exakt selben Typ und Release-Stand wie das Wirtsbetriebssystem.

Die bekanntesten Vertreter dieser Gattung sind OpenVZ/Virtuozzo, Linux Vserver, BSD Jails, OpenSolaris Zones.

Paravirtualisierung

Diese noch recht neue Form stellt einen Kompromiss zwischen SKI und (vollständiger) Virtualisierung dar. Voneinander unabhängige virtuelle Maschinen auf Basis ihres eigenen Betriebssystems greifen über eine bereitgestellte API direkt auf die gemeinsame Hardware zu – gesteuert und kontrolliert durch den Hypervisor.

Der alles entscheidende Punkt dabei: Dieser Ansatz verbirgt die Virtualisierung nicht vor dem Gast, sondern das virtualisierte Betriebssystem „weiß", dass es virtualisiert läuft und mit dem Hypervisor über eine abstrahierte Hardwareschnittstelle kommunizieren muss. Dies erfordert die Anpassung des Gastbetriebssystems.

Das Wirtsystem besteht einzig aus einem spezialisierten Kernel (Hypervisor) sowie einem privilegierten Betriebssystem für Management-Zwecke.

Bekanntester Vertreter dieser Gattung ist Xen™.

1.1.10.3 Intel® VT

Die Intel® Virtualisierungstechnologie (VT) stellt eine wichtige Komponente für Virtualisierung auf Basis von Mainstream-Servern für künftige Desktop- und Servervirtualisierung dar. Mit der hardwareunterstützten Virtualisierung können Softwarevirtualisierungslösungen sehr stark optimiert werden. Daneben erhöht es die Robustheit: Intel® VT er-möglicht VMMs, die Gastsysteme ohne Modifikationen laufen zu lassen, bei gleichzeitiger Erhöhung des Isolationsgrads der virtuellen Maschinen untereinander.

Konkret bieten die Intel® VT Prozessoren Virtualisierungsunterstützung für folgende Hardwarekomponenten:

- CPU
- RAM
- I/O (auf Basis einer IOMMU = I/O Memory Management Unit) (in Vorbereitung)

Erreicht wird die Virtualisierungsunterstützung zum einen durch das Eliminieren der problematischen Instruktionen und zum anderen durch diverse Erweiterungen der Prozessorarchitektur:

- erweiterter Satz an CPU-Instruktionen
- neuer exklusiver Operationsmodus des Prozessors speziell für den VMM/Hypervisor
- Kommunikation (transitions, handoffs) zwischen VMM und Gast über in Hardware implementierte Methoden
- Arbeitsspeicher-Schutz: Der Prozessor reserviert einen exklusiven Speicherbereich für jedes Gastsystem.

Die neue Prozessorengeneration mit Virtualisierungsunterstützung (VMX) schafft eine neue Ebene, die speziell dem Hypervisor vorbehalten ist: VMX Root (vs. VMX Nonroot) (vgl. Abb. 2).

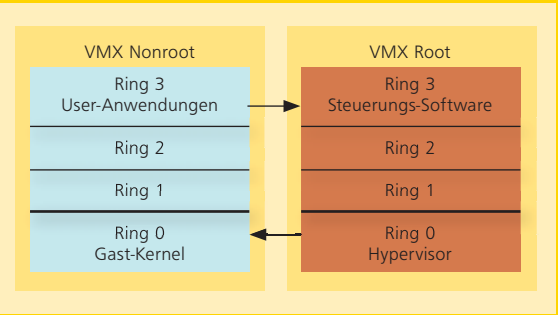


Abb. 2: VMX Root vs. VMX Nonroot

Diese Prozessoren kennen die drei Ringe wie bereits dargestellt für den Nonroot-Bereich, in dem die Gäste laufen; außerdem aber einen parallelen, identisch aufgebauten Root-Bereich für die virtuellen Maschinen. Diese Architektur schafft neue Kontrollstrukturen, um dafür zu sorgen, dass sich Gast und Hypervisor nicht in die Quere kommen und für die Gäste jeweils separate Adressräume geschaffen werden können.

Der Root-Bereich ist für den Hypervisor vorgesehen; sein Verhalten entspricht dem eines normalen Prozessors ohne VT-Komponenten. Die Non-Root Operation stellt eine weitere IA-32-Umgebung zur Verfügung, welche vom Hypervisor kontrolliert wird.

Beide Operationsbereiche unterstützen dabei sämtliche vier Privilegienebenen, so dass Gast-Betriebssysteme in den ursprünglich für sie vorgesehenen Ringen laufen können.

Drei VT-Komponenten sind vorgesehen:

- VT-x: IA-32 Architektur, z. B. Xeon® Prozessoren
- VT-i: Itanium® Architekturen
- VT-d: I/O-Virtualisierung

VT-x definiert zwei neue Transitions: Die Transition von VMX Root Operation zu VMX Nonroot Operation wird als VM entry bezeichnet; eine Transition von Nonroot zu Root Operation als VM exit. Diese Transitions werden von einer eigens neu geschaffenen Virtual Machine Control Structure (VMCS) gesteuert. Diese umfasst einen Gastzustands- und einen Hostzustands-Bereich, welche jeweils Felder enthalten, die die Zustände der einzelnen Areale speichern.

Mit der VMXON-Instruktion wird der VMX-Modus eingeleitet. VM entries (VT-Instruktion VMLAUNCH) laden Prozessorstatus-Informationen aus dem Gastbereich – die Kontrolle ist hiermit an den Gast übergeben. VM exits (VT-Instruktion VMRESUME) speichern Prozessorstatusinformationen in die Gastbereiche und laden danach den Zustand aus dem Hostbereich (vgl. Abb. 3). Mit VMXOFF wird wieder in den Nonroot-Modus zurückgeschaltet und damit die Kontrolle an den VMM zurückgegeben.

Kurz gesagt laufen Gastsysteme im VMX Nonroot-Bereich unverändert und isoliert sowie parallel zum bestens geschützten VMM. Der VMM wiederum kann mittels der neuen Instruktionen sowie der VMCS die Kontrolle an eine virtuelle Maschine übergeben oder ihr wieder entziehen.

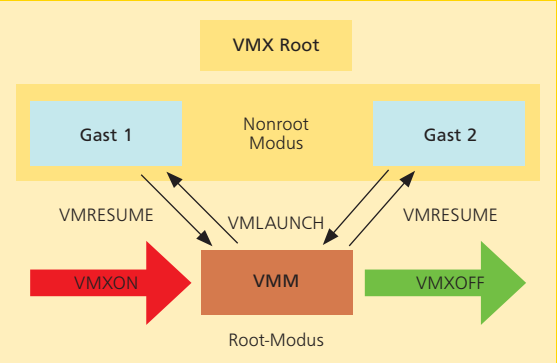


Abb. 3: VMX-Modus

VT-i erweitert Itanium® Prozessoren um die genannten Komponenten sowie darüber hinaus um den Processor Abstraction Layer (PAL). Dieser eröffnet dem Hypervisor eine Schnittstelle, über die er mittels Virtual Processor Descriptor (VPD) z. B. logische Prozessoren definieren kann.

Unter anderem wurde das Prozessorstatus-Register-Bit PSR.vm eingeführt, um einen neuen Ausführungsmodus zu definieren: Der Hypervisor läuft ohne dieses Bit, die Gäste mit gesetztem Bit. Ist dieses gesetzt, lösen privilegierte Prozessorinstruktionen, welche zur Offenlegung des eigentlichen Prozessorstatus führen, Faults aus, die der VMM abfangen und behandeln kann.

VT-d (d für directed I/O) ist dabei die kommende Technologie, welche die vorhandenen VT-Lösungen ergänzt um eine in Hardware implementierte I/O-Gerätevirtualisierung.

Nutzen von VT

Die genannten neuen Ansätze bieten eine Reihe von Vorteilen, in dem sie die Servervirtualisierungstechnik deutlich verbessern:

- sicherer Betrieb durch kleineren, einfacheren Hypervisor
- Kosteneinsparung durch Einsatz „normaler", unveränderter Betriebssysteme als Gäste
- erhöhte Sicherheit durch Isolierung der virtuellen Maschinen bereits in der CPU
- vereinfachte (Weiter-)Entwicklung künftiger Hypervisorsoftware

Wie kann VT genutzt werden?

Folgende Softwarelösungen für Virtualisierung können heute bereits Intel® VT nutzen:

- KVM: Linux-Kernel-Schnittstelle, welche direkt die Intel® VT Schnittstelle anspricht, um unmodifizierte Gastsysteme laufen zu lassen
- Parallels Workstation
- TRANGO real-time embedded hypervisor
- VMware® Workstation 5.5: für 64 Bit-Gäste ist Intel® VT Voraussetzung
- Virtual Iron 3.0, abgeleitet von Xen™
- Xen™ 3.x

Angekündigt sind folgende Softwarepakete:

- Microsoft® Virtual PC 2007 (Beta)
- Microsoft® Virtual Server — VT-Support ist angekündigt für 2007
- VMware® ESX

Autor:

Andrej Radonic ist Vorstand der interSales AG Internet Commerce in Köln. Er befasst sich beruflich seit 12 Jahren mit Internettechnologien, Open-Source-Software, Server Based Computing und Servervirtualisierung. Er ist Autor des Buches Xen 3 (Franzis' Verlag, Oktober 2006), in welchem die quelloffene Virtualisierungslösung sehr ausführlich und praxisorientiert dargestellt wird.

3. Cluster

3.5 Cluster- und Grid-Managementsysteme

Ein Clustersystem besteht vereinfacht aus zwei Arten von Rechnern: einem oder mehreren Servern und mehreren Clients, die über ein schnelles Netzwerk Daten austauschen.

Der prinzipielle Aufbau eines Clustersystems ist meistens asymmetrisch, wobei ein Masterknoten als Verwaltungsrechner für den Cluster fungiert, Aufgaben entgegennimmt und verteilt. Die Zusammenfassung mehrerer Clustersysteme zu einem großen regional verteilten System mit vielen, auch heterogenen Systemen, nennt man Grid. Der Begriff des Grid ist vom englischen Begriff des Power Grid abgeleitet, was Stromnetz bedeutet. Der Hauptgedanke ist die Möglichkeit, immer und überall Rechnerkapazitäten wie Strom aus der Steckdose verfügbar zu machen.

Um eine Ansammlung von Rechnern zu einem Clustersystem zusammenzufügen und die Systeme für den Anwender wie ein einheitliches System wirken zu lassen, wird in vielen Fällen eine komplexe Managementsoftware eingesetzt. Das Hinzufügen und Entfernen von Rechenknoten sowie die Benutzerverwaltung ist dabei nur eine Disziplin. Aufwändige Software- und Konfigurationsupdates können verwaltet und gezielt eingefügt werden. Managementsysteme für den Einsatz auf Cluster- oder Gridsystemen haben das Ziel, die Komplexität der Installation und der Konfiguration zu verstecken.

Viele Anbieter von Cluster-Managementsystemen konzentrieren sich auf die Programmierung von Anwendertools, welche die Bedienung des Gerätes vereinfachen sollen. Meistens läuft diese Software auch nur auf der Hardware des jeweiligen Herstellers und die Nutzung ist durch Lizenzen geregelt.

Mit dem Begriff Cluster- bzw. Grid-Managementsystem werden hier Software-Zusammenstellungen bezeichnet, welche die Einrichtung eines Cluster- oder Gridsystems übernehmen. Diese Einrichtung betrifft viele Bereiche und besteht nicht nur aus der Verwaltung der Benutzer oder der Auslastungsüberwachung. Somit steht der Begriff Cluster- und Grid-Managementsystem für die folgenden Tätigkeiten und Bereiche eines Cluster- oder Gridsystems:

- Installation
- Einrichtung/Konfiguration
- Unterstützung bestimmter Hardware
- Kontrolle/Überwachung des Systems
- Pflege/Aktualisierung (hard- und softwaretechnisch)
- Benutzerverwaltung
- Bereitstellung relevanter Software (parallele Programmierumgebungen, Ressourcenmanagementsysteme etc.)

Vergleichbar sind diese Managementsysteme mit Lösungen für den Client/Server-Betrieb in großen Netzwerken.

Aus anfänglichen Skripten und kleinen Hilfsprogrammen für die Verwaltung der Rechner hat sich eine Reihe mächtiger Werkzeuge entwickelt. Viele dieser Werkzeugpakete basieren auf Open Source, andere auf kommerzieller Basis. Ein Teil dieser Werkzeugpakete zielt stark auf den Einsatz in Rechnernetzwerken, wie sie in Firmen existieren. Diese Pakete beinhalten für den Einsatz bei Cluster- und Gridsystemen bereits viele nützliche Werkzeuge. Darüber hinaus gibt es speziell auf den Cluster- und Gridbereich abgestimmte Werkzeuge.

3.5.1 Hauptfunktionen eines Managementsystems

Managementsysteme gehören zur nächsten Stufe in der Entwicklung der Rechnersysteme. Während es in kleineren Computernetzen mit einigen wenigen Rechnern leicht möglich ist, „Turnschuh-Administration" zu betreiben, würde es in großen Netzwerken mit mehreren hundert oder sogar tausend Rechnern Tage dauern, die gleiche Aufgabe auf diese Weise zu bewältigen. Zu diesen Aufgaben zählen nicht nur die Installation eines neuen Programms, sondern auch komplexere Tätigkeiten, wie zum Beispiel die Statusüberprüfung wichtiger Funktionen. Der Vorteil der gegenüber Gridsystemen räumlichen Nähe der meist in einem Rack montierten Clustersysteme wird durch die sehr große Anzahl an Rechnern zunichte gemacht.

Eine allgemein gültige Aussage über Mindestanforderungen an ein Managementsystem ist nicht möglich. Es hängt vom jeweiligen Einsatzzweck ab, welche Unterstützung und welche Werkzeuge wünschenswert sind. Zum Teil können eigene Skripte und Programme als vollwertiger Bestandteil in die Managementsysteme eingebunden werden. Dadurch ist eine gute Anpassung an die örtlichen Gegebenheiten möglich.

Es lässt sich jedoch eine Maximalanforderungsliste mit allen derzeit für den Einsatz in Cluster- und Gridsystemen nötigen und möglichen Werkzeugen erstellen. Die Betrachtung heutiger auf dem Markt verfügbarer Linux-basierter Systeme ergibt die Gesamtmenge der Funktionen. In nachfolgender Tabelle 1 sind die einzelnen Funktionen der aktuellen Managementsysteme in Hinsicht auf den Betrieb eines Clusters oder Grids aufgenommen:

Managementfunktionen
Monitoringfunktionen
Hardware-Inventur
Konfigurationsverwaltung
Repository
Lights-Out-Management (LOM)
Health-Check
Benutzerverwaltung
Parallele Programmierumgebungen
Ressourcenmanagementsysteme
Unterstützte verteilte Dateisysteme (Clusterdateisysteme)
Weitere Auswahlkriterien
Softwarelizenzierungsmodell
Support
Unterstützte Hardware
Unterstützte Betriebssysteme

Tabelle 1 – Funktionen der Managementsysteme.

3.5.1.1 Managementfunktionen

Dieser Abschnitt befasst sich mit den essentiellen Managementfunktionen einer Cluster- und Grid-Managementlösung. Diese Funktionen dienen dazu, anstehenden Tätigkeiten zu analysieren, evaluieren und eventuell auftretende Probleme zu lösen.

Monitoringfunktionen

Monitoringfunktionen sind bei Systemen für verteiltes Rechnen eine grundlegende Einrichtung. Mit ihrer Hilfe kann der Administrator Daten über die Auslastung, den Speicherverbrauch, die Temperatur und andere Faktoren überwachen, speichern und auswerten und er kann gegebenenfalls eingreifen. Teile dieser Informationen können in vereinfachter Weise dem Benutzer zugänglich gemacht werden.

Dieser ist dadurch in der Lage, den optimalen Ort und Zeitrahmen für seinen Job auszuwählen.

Hardware-Inventur

Hardware-Inventur wird eher in großen Client/Server-Netzwerken benötigt. Die Clients geben Informationen wie Art und Taktfrequenz des Prozessors, Speicherausbau, Netzwerkkonfiguration und Kapazität der Festplatte(n) an den Server weiter. Auf diesen Ergebnissen könnte später die Wahl des Rechnerpools für die Berechnung eines Jobs gründen.

Konfigurationsverwaltung

Die Konfigurationsverwaltung dient dem Wiederherstellen bestimmter Konfigurationen. Bei großen verteilten Systemen kann eine Konfigurationsverwaltung helfen, unterschiedliche Gruppen von Rechnern mit gleicher Software und gleichen Einstellungen auszustatten.

Repository

Server mit Repository-Unterstützung verwalten einen bestimmten Bestand an Softwarepaketen. Diese können zum Beispiel durch Nachfrage bei einem übergeordneten Server aktualisiert werden. Die Verteilung der Software an die Clients übernimmt ein Scheduler zu einem bestimmten Termin. Das Entfernen von Paketen ist umgekehrt auch möglich.

Lights-Out-Management (LOM)

Lights-Out-Management stellt Funktionen wie Hardware-Reset, Reboot und das Auslesen bestimmter Sensoren für Temperatur, Lüfterdrehzahl etc. bereit. Die Softwarefunktionalität muss dabei auf entsprechender Hardware gründen. Ein weit verbreiteter Industriestandard ist IPMI (INTEL® Intelligent Platform Management Interface). Für eine Unterstützung des Hardware-Reset und Reboot muss ein spezieller Out-Of-Band Chip eingesetzt werden. Dieser intelligente Mikrocontroller (Baseboard Management Controller, BMC) bildet das Herzstück des LOM und sorgt für das hardwareseitige „Wohlergehen" des Systems. Er sammelt Informationen und kann auf festgelegte Ereignisse reagieren.

Health-Check

Unter Health-Check versteht man das Überwachen bestimmter Dienste. Am Health-Check-Modul können besondere Dienste registriert werden, welche dann in regelmäßigen Abständen auf Funktion überprüft werden. Der Health-Check kann auch die Überwachung zuvor definierter Eckwerte, wie zum Beispiel Prozessorauslastung übernehmen.

Benutzerverwaltung

Ein zentraler Aspekt großer Cluster- und Gridsysteme ist die Benutzerverwaltung. Verschiedene Benutzer benötigen unterschiedliche Benutzerrechte.

3.5.1.2 Parallele Programmierumgebungen

Das Kernelement eines Cluster- bzw. Gridsystems bildet die parallele Programmierumgebung. Durch Parallelisierung werden Prozesse in Teilstücke aufgeteilt, welche parallel und zeitgleich auf verschiedenen Rechnern ausgeführt werden können. Diese Aufteilung kann durch den Programmierer der Anwendung geschehen, der explizit die Aufteilung eines Prozesses in mehrere Threads unternimmt. Andernfalls erfolgt diese Aufteilung automatisch, so dass kausal unabhängige (nebenläufige) Anwendungen entstehen. Die Aufteilung kann vom Compiler vorgenommen werden.

Der Erfolg der Clustersysteme beruht auf der Entwicklung von Message Passing Bibliotheken (Message Passing Libraries). PVM und MPI sind die erfolgreichsten unter ihnen.

Viele Cluster-Managementsysteme liefern eine bestimmte oder mehrere parallele Programmierumgebungen bereits „out-of-the-box" mit, so dass eine einfache und reibungslose Installation gewährleistet ist.

3.5.1.3 Ressourcenmanagementsysteme

Wie bei der Auslastung eines öffentlichen Verkehrssystems, kostet auch bei einem Cluster bzw. Grid jede nicht genutzte Rechenzeit unnötig Geld. So ist das Ziel der Ressourcenmanagementsysteme (Scheduler/Batch-Queuing-System), eine Auslastung von 100 % zu erreichen.

Das Ressourcenmanagementsystem macht zusammen mit der parallelen Programmierumgebung ein Rechnernetzwerk zu einem Rechencluster.

Unterschieden werden können die Ressourcenmanagementsysteme anhand ihrer Hierarchie und anhand der unterstützten Granularität. Die Hierarchie der meisten Systeme ist linear. Dies bedeutet, dass der Masterknoten als obere Instanz die Aufgabe entgegennimmt, die Aufgabe gegebenenfalls zerteilt und den Rechenknoten zuweist. Bei einer flachen Hierarchie kann jeder Knoten selbst darüber bestimmen, welchem anderen Knoten er eventuell eine Aufgabe übergibt.

Ein Prozess kann aus einem oder mehreren für den Prozessor ausführbaren Teilen, den so genannten Threads, bestehen. Ein Scheduler mit einer bestimmten Granularität hat nun die Möglichkeit, einzelne Threads oder ganze Prozesse auszulagern.

Die drei wichtigsten Ressourcenmanagementsysteme sind:

- OpenPBS
- Sun Grid Engine (SGE)
- Globus

3.5.1.4 Verteilte/Cluster-Dateisysteme

Verteilte Dateisysteme lassen physikalisch verteilte Daten für den einzelnen Rechenknoten wie lokale Daten wirken. Die Benutzer müssen den physikalischen Speicherort der Daten nicht kennen.

Cluster File Systeme erlauben konkurrierende Zugriffe der Rechner aus dem Cluster auf einen gemeinsamen Speicher (Shared Storage). Diese Dateisysteme finden oft Anwendung in High Availability Clustern, wo sie den Zugriff auf den gemeinsamen Speicher managen und kontrollieren. In High Performance Clustern dienen diese Dateisysteme der Aufnahme von großen Datenmengen in sehr kleiner Zeit, wie sie zum Beispiel am CERN, dem weltweit größten Forscherzentrum auf dem Gebiet der Teilchenphysik anfallen.

Sind kurze Zugriffszeiten und eine große Bandbreite gefragt, gibt es auf der Hardwareseite Alternativen. Mit einem Network Attached Storage (NAS) Server lassen sich gute Ergebnisse erreichen. In diesem Fall existiert ein zusätzlicher Server mit Speicheranbindung, welcher für die Vermittlung von Daten zuständig ist. Der nächste Schritt wäre der Einsatz eines Storage Area Network (SAN), bei welchem der Zugriff über das Netzwerk direkt auf Blocklevelebene des Speichers erfolgt. Beide Hardwarelösungen haben den Nachteil des relativ hohen Preises. Günstiger ist hier der Einsatz einer Softwarelösung.

3.5.2 Weitere Unterscheidungsmerkmale

Cluster-Managementsysteme lassen sich des Weiteren auch anhand externer Faktoren unterscheiden. Das Softwarelizenzierungsmodell hat zum Beispiel großen Einfluss auf die Gesamtkosten; auch die Supportkosten beeinflussen die Total Cost of Ownership stark.

3.5.2.1 Softwarelizenzierungsmodell

Unter welcher Lizenz eine Software auf den Markt gebracht wird, unterliegt der Entscheidung des Programmautors bzw. der Softwarefirma. Die Software kann unter einer proprietären oder unter einer freien Lizenz (Open Source) veröffentlicht werden. Im Bereich der Cluster-Managementsysteme existieren verschiedene Lizenzierungsmodelle. Unterscheiden lassen sich die frei verfügbaren und die kostenpflichtigen Managementsysteme, wobei diese hier meist eine Lizenzgebühr pro Knoten, zuzüglich einer weiteren Gebühr für die Grundinstallation erheben.

3.5.2.2 Support

Ein wichtiger Aspekt bei der Anschaffung und beim Betrieb eines Cluster- oder Gridsystems ist die Unterstützung des Herstellers. Eine kleine Anzahl Supercomputer der Top 500 Liste (www.top500.org) der schnellsten Rechner der Welt trägt in der Beschreibung das Stichwort „self-made". Bei diesen Systemen kann von einem engagierten Team von Administratoren ausgegangen werden, welches die Pflege und Verwaltung des Systems ausführt. Nur in seltenen Fällen und bei kleineren Systemen wird der Hardware-Support ebenfalls durch ein internes Team geleistet. In den meisten Fällen ist nicht nur der Hardware-Support, sondern auch die Administration durch ein externes Unternehmen gewährleistet, welches auch die Software für die Anlage verwaltet. Für den Support muss der Entscheider prüfen, welche der beiden Alternativen für die Dauer der Einsatzzeit des Systems günstiger ist.

3.5.2.3 Unterstützte Betriebssysteme (Master/Clients)

Oft ist die Wahl des Managementsystems abhängig von der Unterstützung der Betriebssysteme. Für Managementsysteme in Client-Netzwerken ist diese Unterstützung sehr viel wichtiger als für Cluster- und Gridsysteme. In den meisten Fällen wird das Betriebssystem je nach Anforderung und zu lösender Aufgabe ausgewählt.

3.5.2.4 Unterstützte Hardware

Die besten Softwareeigenschaften einer Cluster- und Grid-Managementlösung nützen nichts, wenn die geforderte Hardware nicht unterstützt wird. Bei Managementlösungen aus einer Hand hat die unterstützte Hardware den entscheidenden Einfluss auf die Realisierung des Systems bzw. auf die Gesamtkosten.

Neben der Architektur spielt das unterstützte Netzwerk eine ebenso wichtige Rolle. Für viele Clusteranwendungen ist die Latenzzeit des Netzwerks der Flaschenhals. Aus diesem Grund sind viele verteilte Rechnersysteme mit einer speziellen Netzwerktechnik ausgestattet, welche niedere Latenzzeiten bietet.

Durch den Schritt der Supercomputertechnologie von SMP- zu Clustersystemen eröffneten sich neben vielen neuen Möglichkeiten auch bestimmte Schwierigkeiten. Während bei einem SMP-System alle Prozessoren über einen schnellen Bus gemeinsam auf den Speicher des Systems zugreifen, muss bei Clustersystemen das Verbindungsnetzwerk mit seinen spezifischen Protokollen diese Aufgabe erledigen.

Die grundlegende Idee heutiger Clustersysteme, ein virtuelles SMP-System zu erschaffen, hat einen entscheidenden Flaschenhals: das Verbindungsnetzwerk. Im Gegensatz zu den SMP-Systemen muss bei den Clustersystemen das Verbindungsnetzwerk die Aufgabe des Busses übernehmen. Der Nachteil der Verbindungsnetzwerke ist der Overhead der Netzwerkprotokolle und die Latenzzeit. Je mehr Rechner zu einem virtuellen SMP-System zusammengeschlossen werden, umso größer ist die Diskrepanz zwischen idealem und realem Zugewinn an Rechenleistung. Diese Diskrepanz wird durch die interne Clusterkommunikation erzeugt. Der entscheidende Faktor ist in diesem Falle nicht die Bandbreite des Verbindungsnetzwerkes, sondern die Latenzzeit.

Aus Gründen der Effektivität wird bei fast allen aktuellen Netzwerktechniken auf eine kleine Latenzzeit geachtet, um die optimale Skalierung (Speedup) der Systeme zu erreichen. Der Durchsatz ist für den optimalen Speedup zweitrangig, jedoch für viele Anwendungen wichtig.

Derzeit aktuelle Netzwerktechniken sind:

- Ethernet (10/100/1000/10.000 MBit/s)
- InfiniBand™
- Myrinet
- Quadrics
- SCI Dolphin

3.5.3 Zusammenfassung

Die kommerziellen Produkte s.cluster mit scVENUS und Scali Manage können in den Bereichen Support und monolithisches Auftreten punkten. Der Preis muss im Endeffekt zu den Anschaffungskosten hinzuaddiert werden und erhöht den „Preis pro Rechenleistung"-Quotienten.

Die nachfolgende Tabelle enthält eine Beschreibung der charakteristischen Merkmale der aktuell am Markt verfügbaren Managementsysteme:

Managementsystem	Kurzbeschreibung/Merkmale
CLUTO	Das flexible CLUTO erlaubt die Anpassung an viele Plattformen und Architekturen und die Verwendung einer breiten Palette an Middlewares und Applikationen.
OSCAR	Die Zusammenstellung der Open Cluster Group hält was sie verspricht: einfaches Aufsetzen und Konfigurieren eines Linuxclusters. OSCAR ist für bestimmte Linux-Distributionen erhältlich.
Rocks Linux	Die Verwaltung der Software und Hardware durch Rolls ist eine elegante Möglichkeit, die Installations- und Konfigurationsarbeit eines Clusters zu vereinfachen. Durch das extrem monolithische Auftreten kann die Installation erschwert oder sogar verhindert werden.
s.cluster/ scVENUS	Die kommerziellen Produkte s.cluster und scVENUS bietet viele Möglichkeiten für die Verwaltung vieler Rechner im Netzwerk. Für die Verwaltung einer großen Anzahl an Rechnern in einem heterogenen Netzwerk ist scVENUS eine gute Lösung für Großunternehmen.
Scali Manage	Scali Manage lässt sich bei transtec als Option erwerben. Durch die Einbindung von Scali MPI Connect in Scali Manage ist eine MPI-Implementierung mit kommerzieller Unterstützung präsent. Zusätzlich bietet Scali Manage mit der Einbindung von PBS Professional™ ein kommerziell unterstütztes Batch-Queueing-System an.

Tabelle 2 – Kurzbeschreibung der evaluierten Cluster- bzw. Grid-Managementsysteme.

Zusammenfassend lassen sich folgende Stichpunkte für die Wahl des Cluster- bzw. Grid-Managementsystems festhalten:

- Je mehr von der Installation und der Konfiguration versteckt wird, desto einfacher wird das Aufsetzen des Clusters, aber desto schwieriger werden Fehlerdiagnose und -behandlung.
- Ein Managementsystem, welches On-Top auf eine Linuxdistribution installiert wird, bietet mehr Flexibilität und Hardwareunabhängigkeit.
- Je nach Anwendungsfall kommt unterschiedliche Anwendungssoftware zum Einsatz. Diese Auswahl ist mitentscheidend für die Wahl des Managementsystems. Die kommerziellen Produkte

verfügen über eine ausgereifte Softwareverwaltung und erleichtern die Installation und Konfiguration der Middleware und Applikationen.

- Die Größe und Struktur des Clusters benötigt unter Umständen eine andere Managementlösung. Beschränkungen oder Präferenzen werden durch Heterogenitätsverwaltung und Architekturunterstützung geschaffen.
- Herstellerseitige Unterstützung für die Clusterinstallation und -administration sowie die Unterstützung für die Einrichtung und Benutzung der Anwendungssoftware sind ein entscheidender Kostenfaktor.
- Zusätzliche Kosten durch den Kauf oder die Miete der Managementlösung müssen mit den Kosten, welche durch persönliche Administration entstehen, verglichen werden. Trotz allem ist der Trend der Zukunft, dass die Kosten für die Applikation in den meisten Fällen mit den Kosten für die Anschaffung der Clusterhardware und Managementlösung gleichziehen.

Durch Beobachtung des Marktes und Einstufung der Cluster- und Gridsysteme sowie durch genaues Studium der einzelnen Merkmale der Managementlösungen lassen sich drei Kategorien mit jeweils einer Lösung zuordnen:

Kategorie/Einsatzzweck des Cluster- bzw. Gridsystems	Optimales Managementsystem
Homogener Cluster bis 64 Knoten	CLUTO
Geschlossene Benutzerverwaltung	
Individuelle Konfiguration gewünscht	
Homogener Cluster mit mehr als 64 Knoten	Scali Manage
Gehobene Benutzerverwaltung	
Kommerzieller Supportanspruch mit Garantieleistung	
Heterogener Cluster mit vielen Knoten (> 64)	s.cluster mit scVENUS
Integration mehrerer Plattformen und Architekturen	
Mögliche Verwaltung weiterer Rechner mit Netzwerk	
Aufwändige Benutzerverwaltung	
Kommerzieller Supportanspruch mit Garantieleistung	

Tabelle 3 – Mögliche Zuordnung eines Cluster- bzw. Grid-Managements zum Einsatzzweck.

5. Festplatten und RAID

5.7 Perpendicular Magnetic Recording

5.7.1 Einleitung

Bei Festplattenspeichern wurde eine faszinierende neue Technologie zur magnetischen Aufzeichnung eingeführt. Magnetische Senkrechtaufzeichnung (Perpendicular Magnetic Recording, PMR) bietet dem Kunden höhere Kapazitäten, verbesserte Zuverlässigkeit und Robustheit sowie äußerst positive Aussichten für zukünftiges Wachstum im Bereich von Kapazität und Leistung. Die Entwicklung der Laufwerke enthielt einen jahrelangen Praxistest, der die Möglichkeiten dieses neuen, fortgeschrittenen Designs erfolgreich demonstrierte. Die Kerntechnologien bestehen aus Köpfen der zweiten Generation mit hinterer Abschirmung und Medien aus hochentwickeltem granularem Kobalt-Chrom-Platin (CoCrPt) Oxid. Gemeinsam optimierte Designs von Köpfen und Medien sowie verbesserte Systemintegration führten zu einer sehr hohen Leistung mit steilen Schreibfeldgradienten, Unempfindlichkeit gegenüber Streufeldern sowie einer exzellenten magnetischen und mechanischen Stabilität.

5.7.1.1 Grundlagen zur Speicherung auf Festplatten

Alle Festplatten speichern die Daten in Form von winzigen Bereichen mit positiver oder negativer magnetischer Polung auf den Oberflächen der Scheiben. Jeder dieser winzigen Bereiche repräsentiert ein Daten-„Bit". Die Bits werden eng aneinander auf kreisförmigen „Spuren" der rotierenden Plattenoberfläche geschrieben. Die Oberflächen der Platten enthalten viele dieser konzentrischen Spuren. Jede Spur enthält Millionen von Bits und jede Plattenoberfläche enthält viele Zehntausende von Spuren. Die Gesamtspeicherkapazität einer Festplatte hängt direkt davon ab, wie klein sich der zur Darstellung eines Datenbits erforderliche Bereich machen lässt: Je kleiner die Bits, desto größer ist die Kapazität.

5.7.1.2 Aufzeichnungsdichte, Technologiewachstum, thermischer Grenzwert

Das Produkt aus Bits pro Zoll entlang der Spur und radial vorhandenen Spuren pro Zoll auf der Plattenfläche ist die Aufzeichnungsdichte in Bit pro Quadratzoll. Die Wachstumsrate für die Aufzeichnungsdichte ist ein häufig genanntes Maß für den Fortschritt dieser Technologie. In den letzten Jahren hat sich die Wachstumsrate verlangsamt, da es bei magnetischer Aufzeichnung einen fundamentalen Grenzwert gibt. Dieser Grenzwert steht in Zusammenhang mit der Tatsache,

dass das magnetische Material auf der Plattenoberfläche notwendigerweise aus kleinen Körnern besteht. Aufgrund der Zufälligkeit von Kornform und -größe muss jedes auf die Platte geschriebene Bit ca. 100 Körner abdecken, um eine zuverlässige Speicherung der Daten sicherzustellen. Unglücklicherweise gibt es für die Korngröße einen unteren Grenzwert. Unterhalb dieses Grenzwertes besteht das Risiko, dass die Magnetisierung sich durch Anregung durch die überall in der Umgebung vorhandene thermische Energie spontan umkehrt, selbst bei Raumtemperatur.

5.7.1.3 Perpendicular Recording Technologie

Senkrechtaufzeichnung verändert diesen „thermischen" Grenzwert und ermöglicht weitere Fortschritte bei der Aufzeichnungsdichte. Bei konventioneller magnetischer Aufzeichnung in Längsrichtung (Longitudinal Magnetic Recording, LMR) ist die Magnetisierung der Bits entlang der Spuren ausgerichtet. Bei Senkrechtaufzeichnung zeigen die „magnetischen Bits" nach oben oder unten senkrecht zur Plattenoberfläche. Abbildung 1 stellt die Konfiguration von Aufzeichnungsmedien, Schreibköpfen und Leseköpfen für Aufzeichnungssysteme in Längsrichtung und senkrechter Ausrichtung gegenüber.

Das einzigartige Merkmal beim System mit senkrechter Ausrichtung ist die „weiche magnetische Unterschicht" innerhalb der Platte. Diese Unterschicht besitzt eine sehr gute Leitfähigkeit für Magnetfluss. Wenn der Schreibkopf aufgeladen wird, konzentriert sich der Fluss unterhalb der kleinen Polspitze und erzeugt ein intensives Magnetfeld im kleinen Spalt zwischen der Polspitze und der weichen Unterschicht. Die Aufzeichnungsschicht, in der die Daten gespeichert werden, befindet sich direkt an der Stelle in diesem Spalt, an der das Feld am stärksten ist. Stärkere Felder ermöglichen die Verwendung von Medien mit „höherer Koerzitivität". Solche Medien erfordern stärkere Felder zum Setzen der Magnetisierung; danach ist die Magnetisierung dadurch allerdings umso stabiler.

Das Vorhandensein der weichen Unterschicht stärkt darüber hinaus die Lesesignale und hilft bei der Verringerung der Interferenzen von benachbarten Spuren. Obgleich am Lesekopf selbst keine großen Veränderungen erforderlich sind, sind die vom Kopf abgegebenen Wellensignale völlig unterschiedlich und erfordern neue Signalverarbeitungstechniken, um den größten Nutzen zu erzielen.

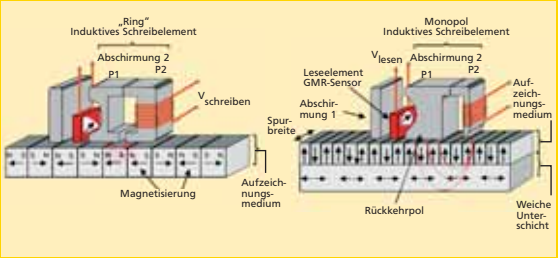


Abb. 1: Diagrammdarstellung für Längsaufzeichnung (links) und Senkrechtaufzeichnung (rechts).

5.7.1.4 Perpendicular Recording Technologie der zweiten Generation

Abbildung 1 zeigt, was man als Technologie der ersten Generation bezeichnen kann. Obgleich sich diese Technologie als vorteilhaft und zuverlässig erwiesen hat, lassen sich durch weitere Verfeinerungen an Köpfen, Medien und Elektronik weitere Fortschritte erzielen. Dies ist die Technologie der zweiten Generation (z. B. In der Hitachi Travelstar® 5K160). Die Technologie der zweiten Generation enthält Änderungen am Schreibkopf, am Aufzeichnungsmedium und an der Lese-/Schreibelektronik.

Der Schreibkopf wird so verändert, dass er eine hintere Abschirmung erhält, die nahe an der hinteren Kante der Polspitze platziert ist, an der die Daten aufgezeichnet werden, wie in Abbildung 2 gezeigt. Dies kann sich geringfügig auf die Feldstärke auswirken. Ein großer Vorteil liegt allerdings darin, dass die Felder sehr schnell abfallen, wenn sich das Medium von der Stelle unterhalb der Polspitze unter die Abschirmung bewegt. Dieses schnelle Abfallen des Feldes bedeutet, dass die geschriebenen Bits viel schärfer definiert werden können.

Zur einfacheren Implementierung waren Medien der ersten Generation aus einer einzelnen homogenen Schicht aufgebaut. Es ist jedoch sehr vorteilhaft, die Eigenschaften im Verlauf des Medienquerschnitts unterschiedlich anzupassen. Diese Eigenschaften sind das magnetische Moment (Magnetisierung pro Volumeneinheit), die Anisotropie (die Bereitschaft, mit der sich die Magnetisierung in einer bestimmten Richtung ausrichtet) und der Austausch (der Grad der atomaren Kopplung zwischen benachbarten Körnern, die bewirkt, dass deren Magnetisierung dazu neigt, in die gleiche Richtung zu zeigen). Diese magnetischen Eigenschaften ergeben sich durch das komplexe Zusammenspiel der verwendeten Materialien und der Bedingungen bei der Herstellung.

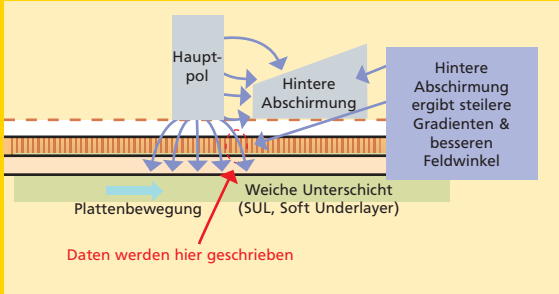


Abb. 2: Querschnittsdarstellung eines Kopfes mit „hinterer Abschirmung" – eine dünne magnetische Abschirmung, die in der Nähe der hinteren Kante des Schreibpols platziert ist.

5.7.2 Kopftechnologien

Bei einem konventionellen LMR-Kopf wird das Magnetfeld zum Schreiben an einem dünnen, nichtmagnetischen Spalt im Ringkopf erzeugt, und seine Längskomponente ist stärker ausgeprägt als seine Vertikalkomponente. Bei PMR ist die Magnetisierung des Mediums nach oben und unten ausgerichtet. Um einen effizienten Schreibvorgang zu erreichen, muss ein PMR-Schreibkopf Felder erzeugen, deren senkrechte Komponenten stärker ausgeprägt sind als die Längskomponenten. Wie in Abb. 1 gezeigt, bietet ein PMR-Kopf mit einem Einzelpol in Kombination mit einer weichen Unterschicht (SUL, Soft Underlayer) ein starkes senkrechtes Schreibfeld mit stark reduzierter Längskomponente. Das Feld von eines PMR-Schreibkopfs wird nicht am Spalt erzeugt, sondern an der Poloberfläche, und von der SUL aufgefangen. Abbildung 3 zeigt, dass die Kanten eines rechteckigen Pols die benachbarten Spuren schneiden, wenn der Kopf in schrägem Winkel zum Spurverlauf betrieben wird. Bei modernen Laufwerken weist der Kopf eine Schräglage in Bezug zur Spurrichtung auf, wenn er sich über den inneren oder äußeren Spuren befindet. Die Herstellung eines schmalen trapezförmigen Pols mit einem gut gesteuerten Schrägwinkel ist eine wesentliche Voraussetzung dafür, das Löschen von Daten auf benachbarten Spuren durch die Poloberfläche zu verhindern.

Der PMR-Kopf der zweiten Generation mit hinterer Abschirmung von Hitachi verfügt über eine genaue Steuerung der Abschirmungsdicke und den Spalt zur hinteren Polkante, um ein exaktes Gleichgewicht der Interaktion zwischen der hinteren Abschirmung und dem Hauptpol zu erreichen. Die genaue Anpassung von Medien an Köpfe mit hinterer Abschirmung ist von großer Bedeutung, um die Vorteile durch hohen Feldgradienten und optimalen Feldwinkel zu nutzen und um sie an die geänderte Feldstärke anzupassen. Beim Schreiben auf Medien mit passenden Charakteristika schreiben Köpfe mit hinterer Abschirmung schärfer definierte Bits. Das Ergebnis davon ist, dass das Laufwerk über eine niedrigere Bitfehlerrate verfügt und daher eine höhere Zuverlässigkeit aufweist.

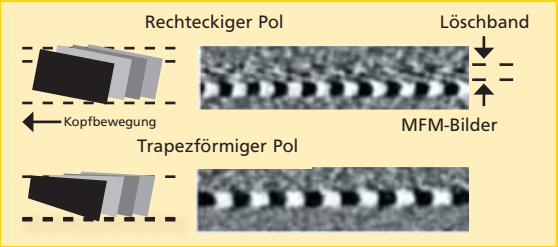


Abb. 3: Trapezförmige Polform bei einem PMR-Schreibkopf zur Vermeidung von Datenlöschung in benachbarten Spuren, wenn der Kopf sich in Schräglage zum Spurverlauf befindet, z. B. im Innen- oder Außenbereich.

Neben der Optimierung des Kopfdesigns für die Aufzeichnungsleistung gibt es weitere Herausforderungen, die nur bei senkrechten Köpfen auftreten, z. B. „Pollöschung" und „Löschung durch Störfelder". Beide Probleme können zu Datenkorruption führen, wenn keine geeigneten Maßnahmen ergriffen werden. Die Pollöschung bezieht sich auf das Phänomen, dass der Schreibpol weiter Magnetfelder erzeugt, selbst wenn der Schreibstrom am Ende des Schreibzyklus auf Null gesetzt wird. Das kann dazu führen, dass Daten unabsichtlich gelöscht werden. Dieses Phänomen ergibt sich aus den extrem kleinen Abmessungen und den ferromagnetischen Eigenschaften der senkrechten Schreibpole. Besondere Anstrengungen bei der Auswahl und Verarbeitung von magnetischen Dünnfilmmaterialien sind erforderlich, um einen Schreibkopf ohne Pollöschung sicherzustellen.

Eines der ersten Probleme mit PMR war eine inhärent höhere Empfindlichkeit für externe Magnetfelder im Vergleich zur Technologie mit Längsaufzeichnung. Die erhöhte Empfindlichkeit für Störfelder resultiert aus der Interaktion zwischen dem Aufzeichnungskopf und der SUL. Diese externen Felder sind ein besonderes Problem für mobile Produkte, wo beispielsweise ein magnetisches Armband bei einem Laptop leicht bis in eine Entfernung von ein paar Zentimetern zur Festplatte gelangen kann. Ohne spezielles Kopfdesign können die externen Felder die Schreib- und Lesesignale stark verzerren und Fehlerereignisse hervorrufen. Bei manchen extremen Fällen kann das externe Feld sogar eine irreversible Löschung von Daten bewirken.

Durch Sorgfalt beim Design von Köpfen und Medien ist es Hitachi gelungen, die Resistenz gegenüber externen Feldern auf ein Niveau zu bringen, das dem bei Laufwerken mit Längsaufzeichnung gleichkommt oder dieses sogar übertrifft.

5.7.3 Medientechnologie

Da Medien für Längsaufzeichnung sich dem unteren Grenzwert für die thermisch stabile Bitgröße nähern, war die Industrie motiviert, die historisch komplexen Probleme zu lösen, die sich aus der Herstellung von Medien für die Senkrechtaufzeichnung ergeben. Die grundlegend entwickelte Medienstruktur für Produkte im Jahr 2006 besteht aus einer Art „granularem" Medium, zusammengesetzt aus magnetischen Legierungen aus Kobalt, Chrom und Platin (CoCrPt) und einem Oxidtrennmittel für die Korngrenzen, wie in Abbildung 4 gezeigt. Durch Verwendung einer Hitachi-eigenen Legierungskombination und eines Verfahrens zur Schichtenaufbringung unterscheiden sich die Aufzeichnungseigenschaften im Verlauf des Medienquerschnitts. Damit wird optimaler Rauschabstand bei gleichzeitig hervorragenden Schreibeigenschaften und hoher mechanischer Qualität erreicht. Medien und Köpfe wurden gemeinsam entwickelt, um Vorteile beim Design von Schreibköpfen mit hinteren Abschirmungen, den weichen magnetischen Unterschichten der Medien und den harten magnetischen Schichten der Medien zu nutzen.

Bevor die Produktentwicklung wirklich begann, wiesen Medien für Senkrechtaufzeichnung eine geringere mechanische Zuverlässigkeit auf als Medien für die Längsaufzeichnung. Gemeinsame und koordinierte Anstrengungen der umfangreichen Ressourcen von Hitachi waren erforderlich, um diese Probleme zu verstehen und die Zuverlässigkeit auf das Niveau anzuheben, das den Qualitätsstandards von Hitachi entspricht, und dabei noch die magnetische Leistung zu verbessern und für die Produktion erforderliche Zykluszeiten und Ausbeuteraten zu erreichen. Bei der Entwicklung von Medien für Senkrechtaufzeichnung waren ein Umdenken in Bezug auf Zuverlässigkeitskriterien sowie Vorsorgemaßnahmen zur Begegnung potenzieller Fehlermechanismen erforderlich, um höchste Korrosionsbeständigkeit und mechanische Robustheit zu erreichen. Darüber hinaus hat die Fertigung von PMR-Medientechnologie neue Ansätze bei der Sputter-Ausrüstung mit mehr Abscheidestationen, höherer Verarbeitungsleistung und höherem Durchsatz erforderlich gemacht.

Abb. 4: Aufbau von Hitachi-Medien zur Senkrechtaufzeichnung.

5.7.4 Lese-/Schreibelektronik

Die Signale aus einem System für Senkrechtaufzeichnung unterscheiden sich grundlegend von denen aus einem konventionellen System zur Längsaufzeichnung (Abbildung 5). Jede Frequenzkomponente erfährt eine Phasenverschiebung um 90 Grad (entsprechend der 90-Grad-Rotation der Magnetisierung von der Längsaufzeichnung zur Senkrechtaufzeichnung). Dadurch wird das Erscheinungsbild der Wellenformen völlig verändert. Die Signalverarbeitung in der Lese-/Schreibelektronik muss zur Verarbeitung dieser Wellenformen angepasst werden. Zusätzlich zur Phasenverschiebung gibt es auch eine viel größere Signalstärke im Niederfrequenzbereich.

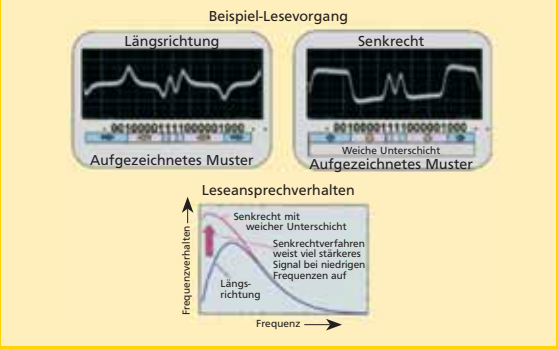


Abb. 5: Die Wellenformen beim Lesen von Längs- und Senkrechtaufzeichnung unterscheiden sich grundlegend. Die Signalverarbeitung im Lese-/Schreibkanal muss an diese neuen Wellenformen angepasst werden und auch einen Teil der zusätzlichen Signalenergie bei niedrigen Frequenzen berücksichtigen.

Quelle: *Hitachi Global Storage Technologies*