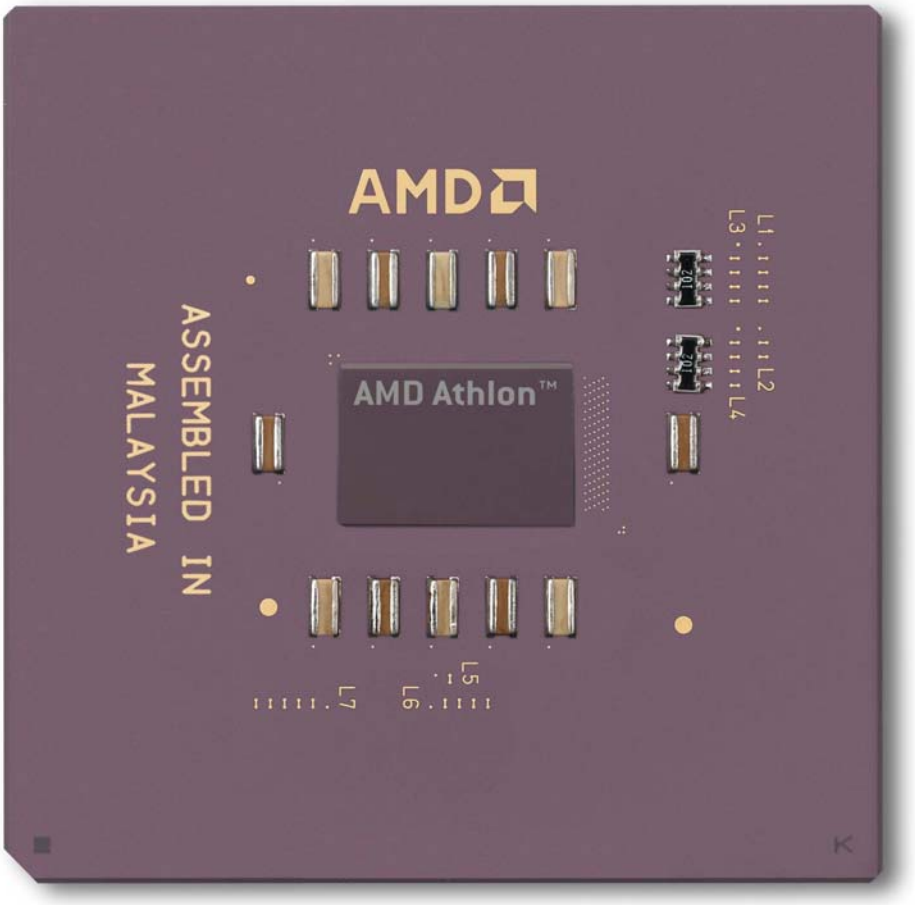
PROZESSORARCHITEKTUR

AM BEISPIEL DES AMD ATHLON

AUSGEARBEITET VON ALEXANDER TABAKOFF



Betreuender Lehrer: Prof. Wolfgang Schinwald

VERÖFFENTLICHT AM 26.2.2001

INHALTSVERZEICHNIS:

1 HISTORISCHE / ALLGEMEINE EINFÜHRUNG

1.1Die Anwendungsbereiche von Prozessoren

Prozessoren haben heute verschiedenste Anwendungsbereiche. Sie werden in Autos, Set Top Boxen, Spielekonsolen, Handys, Taschenrechnern, PCs usw. verwendet. Dabei macht der Marktanteil der PC Prozessoren nur rund 2%[1] aus. Trotz dieser vergleichsweise geringen Produktion genießen PC Prozessoren einen bedeutend höheren Bekanntheitsgrad. Fast jeder kennt PC Prozessoren wie den Intel Pentium oder den AMD Athlon, doch wie viele haben schon von einem IBM PPC 440 oder einem Toshiba TMPR1940CY gehört? Man kann nun die Prozessoren grundsätzlich nach Embedded oder nicht Embedded einteilen, wobei danach auch eine weitere Einteilung in Architekturen zu geschehen hat. Im Bereich der Computerprozessoren sind die wichtigsten Architekturen: x86 (bei IBM AT und kompatibel), PPC (Apple ab PowerMac, IBM Mainframes der S/390 und RS/6000 Serie), MIPS (Silicon Graphics mit Octane,..), SPARK (Sun Microsystems) und Alpha (Compaq und andere). Hier ist die x86 Sparte am Absatz gemessen die größte. Auch in der x86 Architektur gibt es verschiedene Hersteller. Die Vertreter sind Via (mit Cyrix Reihe), Transmeta (Crusoe), Intel (Pentium/Xeon/Celeron Reihen) und AMD (K6/Athlon/Duron Reihen). Doch bevor ich mich mit dem Athlon näher beschäftige, möchte ich noch auf die Historie der x86 Prozessoren eingehen.

1.2Der erste Prozessor

Der erste Prozessor in der Form, in der wir ihn heute haben, war der Intel

1 Vergleiche Andreas Stiller: Prozessorgeflüster. Von Acherbahnen und Karussellfahrten c´t 22/2000, Seite 52

4004. Er wurde 1971 auf den Markt gebracht, besaß 2 300 Transistoren und taktete mit 108 kHz. Er konnte 60 000 Operationen pro Sekunde ausführen, lächerlich im Vergleich zu heutigen Prozessoren, aber damals ein Meilenstein. Er wurde als der erste Computer auf einem Chip bezeichnet, da seine Vorgänger noch alle mit Röhren arbeiteten und ganze Zimmer ausfüllten.

1.3Die Entwicklung bis zum 586

Von diesem Zeitpunkt wurde konsequent weiterentwickelt: Intel produzierte den 8080 und den 8086. Der 8086 spielt eine wichtige Rolle in der weiteren Entwicklung, da alle Prozessoren, die nach diesem von Intel produziert wurden, mit ihm abwärts kompatibel sind.
In den folgenden Jahren kamen 8088, 80286, 80386 und der 80486 und auch immer abgespeckte Einsteiger CPUs (Kürzel SX) auf den Markt.
Doch auch Intel hatte Konkurrenz: Firmen wie Motorola, AMD, Cyrix/IBM, Nexgen und Siemens versuchten immer wieder die CPUs des Marktführers Intel zu imitieren und zu modifizieren, um unter anderem Namen, aber meist um einiges billiger Prozessoren auf den Markt zu werfen. Die Konkurrenten konnten aber nie den Marktführer unter Druck setzen und Intel hatte immer die Leistungskrone inne. Es gab zwar immer wieder leistungsfähigere Prozessoren (z.b.: Zilog Z80, Motorola 68000, Nexgen Nx586) die sich aber nie durchsetzen konnten. Auch die Firma AMD, sie wurde im selben Jahr wie Intel gegründet, stellte Intel kompatible CPUs her, aber konnte nie so recht mit dem Tempo, das Intel vorgab, mithalten. Neuerungen wie der 486 hatten teilweise erst 3 Jahre Verspätung, bis sie auf der AMD Plattform erhältlich waren. Doch mit der Ära des Pentium konnte AMD aufholen: Der Konzern kaufte den Hersteller Nexgen und integrierte das Know-How in den AMD

K6.Dieser war in 16Bit Anwendungen und bei Programmen die überwiegend Integer[2] Berechnungen durchführten, schneller als der Pentium MMX von Intel. Intel reagiert mit dem Pentium II, doch der ist Anfangs noch zu teuer. Als Reaktion auf den Druck von AMD senkt Intel die Preise des Pentium II massiv, und diese Prozessorplattform beginnt sich bei Consumer PCs durchzusetzen. AMD ließ aber nicht lang auf sich warten und konterte mit dem K6-2 3Dnow!, der die Neuentwicklung 3Dnow![2] enthielt, die bei speziell optimierten Anwendungen beeindruckende Performance lieferte. Doch Intel reagierte mit schnelleren Pentium IIs und schließlich mit dem Pentium III und seiner neuen SIMD[2] Erweiterung ISSE [2]. AMD entwickelte den K6 nochmals zum K6-3 3Dnow! weiter, dieser hatte einen integrierten L2 Cache [2], er war aber deshalb sehr teuer und konnte nicht mit dem Pentium III von Intel konkurrieren. AMD arbeitete aber an ihm nicht mehr intensiv weiter (erst viel später kamen noch der K6-2/3+ für den mobilen Einsatz dazu) und konzentrierte sich voll auf sein neuestes Projekt: den AMD Athlon. Als erste Benchmarks und einige Grundzüge des Designs bekannt wurden, horchten die Experten auf, und als er auf den Markt kam, hatte es AMD erstmals geschafft: Sie waren dem weltgrößten PC-Prozessor Hersteller Intel erstmals mit ihrem neuen Design in allen Bereichen überlegen. Mit 1,3% und 40,3%[3] mehr Leistung als ein gleichgetakteter Pentium III zu einem deutlich niedrigeren Preis konnte sich AMD schmücken.

2 Siehe Glossar

3 Vergleiche Daniel Wolff: Intels schlimmster Alptraum. Wachablösung bei den PC-Prozessoren: Der neue Athlon (Code-Name K7) von AMD schlägt im Chip-Testcenter den gleich schnell getakteten Pentium-III in allen Benchmarks, CHIP 9/1999, Seite 132,

1.4 Der AMD Athlon und der Pentium III – Entwicklungsgeschichte

Eine neue Prozessorgeneration hat immer eine lange Entwicklungsgeschichte hinter sich. Aufgrund der Komplexität moderner PC Prozessoren sind eine ausgedehnte Entwicklungsphase und langwierige Tests vonnöten. Die Designer müssen ja auch über 30 Millionen Transistoren in die richtige Verbindung zueinander bringen, dabei sind Fehler aber unvermeidlich (dazu aber später). Man darf aber bei Prozessoren nicht davon ausgehen, dass sie während ihrer Marktpräsenz nicht verändert werden. So ist zum Beispiel zwischen einem Athlon 550, einem Athlon 800 und einem 1,1 GHz Athlon weit mehr Unterschied als nur die MHz Anzahl. Es gibt Unterschiede im Fertigungsprozess, in der Bauform, in der Größe der Caches und im Verbrauch. Nachfolgend ist eine Tabelle zu sehen, auf der alle Athlon Modelle seit dem Debüt im August 1999 und ihre wichtigsten Features bzw. Neuerungen aufgeführt sind[4].

(Zu dieser Tabelle ist noch anzumerken: Bei den Zellen die nicht ausgefüllt sind konnte ich leider die dementsprechende Angabe nicht finden.)

	Takt (in MHz)	***Multiplikator***	***L2 (in KB)***	***Taktverhältniss zu L2***	***Leistung (in Watt)***	***Strukturbreite***	***Steckplatz***	***Core Spannung***	***Modellbezeichnung***
Modell 1	500	5	512	1:2	42	0,25 µm	Slot A	1,6 V	K7 500 MTR51B C
	550	5,5	512	1:2	46	0,25 µm	Slot A	1,6 V	K7 550 MTR51B C
	600	6	512	1:2	50	0,25 µm	Slot A	1,6 V	K7 600 MTR51B C
	650	6,5	512	1:2	54	0,25 µm	Slot A	1,6 V	K7 650 MTR51B C

4 Diese Tabelle wurde aus folgenden Quellen erstellt: c´t 14/2000, Seite 107; AMD Datenblätter 21016 und 23792; Stand ist der 10.12.2000

PROZESSORARCHITEKTUR

	Takt (in MHz)	*Multiplikator*	*L2 (in KB)*	*Taktverhältniss zu L2*	*Leistung (in Watt)*	*Strukturbreite*	*Steckplatz*	*Core Spannung*	*Modellbezeichnung*
	700	7	512	1:2	50	0,25 µm	Slot A	1,6 V	K7 700 MTR51B C
Modell 2	550	5,5	512	1:2	31	0,18 µm	Slot A	1,6	K7 550 MTR51B A
	600	6	512	1:2	34	0,18 µm	Slot A	1,6	K7 600 MTR51B A
	650	6,5	512	1:2	36	0,18 µm	Slot A	1,6	K7 650 MTR51B A
	700	7	512	1:2	39	0,18 µm	Slot A	1,6	K7 700 MTR51B A
	750	7,5	512	2:5	40	0,18 µm	Slot A	1,6	K7 550 MTR52B A
	800	8	512	2:5	48	0,18 µm	Slot A	1,7	K7 550 MPR52B A
	850	8,5	512	2:5	50	0,18 µm	Slot A	1,7	K7 550 MPR52B A
	900	9	512		60	0,18 µm	Slot A	1,8	K7 550 MNR52B A
	950	9,5	512			0,18 µm	Slot A		K7 950 MNR52B A
	1000	10	512	1:3	65	0,18 µm	Slot A	1,8	K7 550 MNR52B A
Modell 4	650	6,5	256	1:1		0,18 µm	Slot A		A 0650 MPR24B A
	700	7	256	1:1		0,18 µm	Slot A		A 0700 MPR24B A
	750	7,5	256	1:1		0,18 µm	Slot A		A 0750 MPR24B A
	800	8	256	1:1		0,18 µm	Slot A		A 0800 MPR24B A
	850	8,5	256	1:1		0,18 µm	Slot A		A 0850 MPR24B A
	900	9	256	1:1		0,18 µm	Slot A		A 0900 MMR24B A
	950	9,5	256	1:1		0,18 µm	Slot A		A 0950 MMR24B A
	1000	10	256	1:1		0,18 µm	Slot A		A 1000 MMR24B A
Thunderbird	650	6,5	256	1:1	36,1	0,18 µm	**Socket A**	1,7	A 0650 APT 3 B
	700	7	256	1:1	38,3	0,18 µm	**Socket A**	1,7	A 0700 APT 3 B
	750	7,5	256	1:1	40,4	0,18 µm	**Socket A**	1,7	A 0750 APT 3 B
	800	8	256	1:1	42,6	0,18 µm	**Socket A**	1,7	A 0800 APT 3 B
	850	8,5	256	1:1	44,8	0,18 µm	**Socket A**	1,7	A 0850 APT 3 B
	900	9	256	1:1	49,7	0,18 µm	**Socket A**	1,75	A 0900 AMT 3 B
	950	9,5	256	11	52	0,18 µm	**Socket A**	1,75	A 0950 AMT 3 B
	1000	10	256	1:1	54,3	0,18 µm	**Socket A**	1,75	A 1000 AMT 3 B
	1100	11	256	1:1		0,18 µm	**Socket A**	1,75	
	1200	12	256	1:1		0,18 µm	**Socket A**	1,75	

Natürlich gäbe es auch für den Pentium 3 Kern eine dementsprechende Tabelle, doch da diese Architektur wesentlich älter ist (Debüt 1995 im Pentium Pro) als die des AMD Athlon (infolgedessen gibt es noch viel mehr Typen) und weil es nicht zum eigentlichen Thema der FBA zählt, wurde diese weggelassen[5].

Die verschiedenen Typen sind auch optisch zu unterscheiden. So wie auf Abb. 1 sieht der AMD Athlon Modell 1, 2 und 4 aus.

Abb. 1

Äußerlich sind zwischen den Modellen mit Ausnahme der Seriennummer keine Unterschiede zu erkennen, beim Öffnen der Plastikverkleidung[6] offenbaren sich aber mehr oder minder deutlich die Unterschiede. Bei Modell 1 und 2 sitzen neben dem Prozessor selbst noch die 512 KB großen Cache-Module auf dem Carrier Modul (deutlich in Abb. 2 zu sehen).

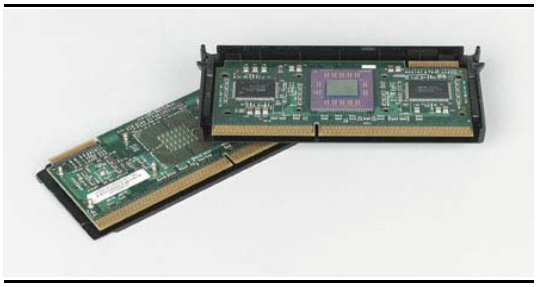


Abb. 2

5 Nachzulesen ist die Tabelle zum Beispiel unter c´t 14/2000, Seite 101 f

6 Vorsicht beim Öffnen! Anleitung z.B. unter CHIP, April 2000, Seite 124 f

In Modell 4 in der Version mit Slot A, in der der L2-Cache im Prozessor integriert ist, sind die zwei großen Speichermodule nicht mehr zu erkennen, wie in Abb. 3 zu sehen ist.

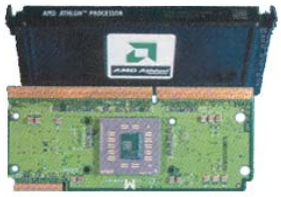


Abb. 3

Diese Version des Thunderbird, so der Codename, ist eher selten anzutreffen und außerdem nicht mit allen Slot A Mainboards kompatibel, da er ein anderes Businterface besitzt.

Der Thunderbird für den Sockel A war die ursprünglich von AMD geplante Version. Er benötigt einen neuen Steckplatz (eben den Sockel A) und besitzt einen 256Kb großen L2 Cache, der On Die[7] sitzt, 64bit breit an den Prozessor angebunden ist und mit voller Prozessor-Taktgeschwindigkeit angesteuert wird. Optisch sind natürlich aufgrund des Wechsels auf den neuen Steckplatz und den Verzicht auf das Carriermodul starke Unterschiede zu erkennen, Abb. 4 zeigt eben diese.

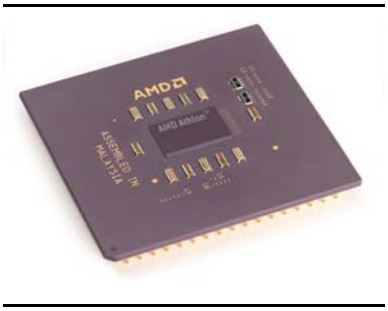


Abb. 4

Der Pentium 3 hat eine längere Entwicklungsgeschichte hinter sich. Sein

7 Siehe Glossar

Design feierte im Penitum Pro Debüt und ist auch im neuesten P3 noch vorhanden. Gestartet ist das Design 1995 mit 150 MHz, heute sind 1Ghz möglich, außerdem hat Intel den neueren Modellen noch zusätzliche Features beigepackt. Der Pentium Pro war der erste Intel Prozessor, der die x86 Befehle nur mehr emulierte. Seine interne Architektur war RISC[8] artig aufgebaut. Er war aber rein für 32Bit Anwendungen ausgelegt, weshalb er nur im Serverbereich Fuß fassen konnte. Sein Nachfolger war der Pentium II, der dann endlich auch im Consumer Bereich erfolgreicher war. Er hatte 2 MMX Einheiten, größere Caches- die aber nur 2:1 getaktet waren- und war bis zu 300 MHz schnell. Er wurde im damals schon veralteten 0,35µm Prozess gefertigt, was bei seinem Nachfolger geändert wurde, denn durch die geringere Strukturbreite des Pentium II Deschutes - so der Codename des Nachfolgers- und einige andere Veränderungen konnte Intel die Taktrate bis auf 450 MHz steigern. Dann folgte der Pentium III mit einer SIMD Erweiterung namens ISSE[9] und einer sehr umstrittenen Seriennummer, die auslesbar war und ermöglichen sollte, einen Prozessor eindeutig zu identifizieren. Die bis jetzt letzte Entwicklungsphase erfährt der Pentium III unter dem Codenamen Coppermine. Anders als der Name vermuten lässt, ist er nicht in einem Kupferprozess gefertigt, sondern nach wie vor in Aluminium, jedoch mit nur mehr 0,18µm Strukturbreite. Als wichtigste Neuerung wird der L2 Cache, dessen Größe 256Kb beträgt, auf dem Die integriert, wodurch auch das Carriermodul überflüssig wird und der Prozessor in einen neuen Steckplatz (FC-PGA)gesteckt werden kann, wobei parallel dazu auch eine Slot Variante erhältlich ist. Abb. 5 zeigt die wichtigsten optischen Unterschiede der P6 Generation und ihrer einzelnen Vertreter. Die Mobilversionen des Pentium 2 und 3 wurden,

8 Siehe Glossar

9 Siehe Glossar

genauso wie die High end Server Prozessoren weggelassen.

Abb. 5

Von links nach rechts sind zu erkennen: Pentium Pro, Pentium II, Pentium III (Katmai), Pentium III (Coppermine).

2 GRUNDLEGENDE DINGE ZUR PROZESSORARCHITEKTUR UND DEM BAU VON PROZESSOREN

2.1 Physikalisch

2.1.1 Der Aufbau eines Transistors

Der Metalloxid-Feldeffekt-Transistor (kurz: MOSFET) ist der wichtigste Bestandteil eines Prozessors, deshalb sollte sein Aufbau genau betrachtet werden. Der MOSFET besteht in seiner Grundform aus sehr einfachen Komponenten. „Zwischen zwei stark dotierten Elektroden liegt ein Kanal der entgegengesetzten Dotierung - bei n-Kanal-MOSFETS liegt also zwischen den n-dotierten Source- und Drain-Elektroden ein p-dotierter Kanal. Über diesem Kanal und durch eine Oxid-Schicht elektrisch davon isoliert befindet sich eine dritte, die so genannte Gate-Elektrode.[10]“Zu sehen ist dies auf Abb. 6.

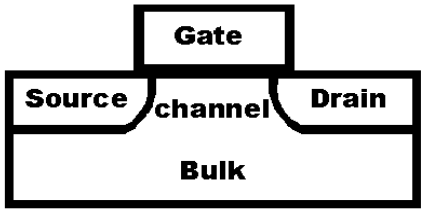


Abb. 6

Nun zur Funktionsweise dieses Transistors. „ Liegt an der Gate-Elektrode eine hinreichend große positive Spannung an, werden Elekronen aus dem p-dotierten Halbleiter angezogen; unter dem Oxid bildet sich ein dünner, leitender Kanal aus. (...) Die Leitfähigkeit des Kanalgebietes (channel) wird durch das auf die Landungsträger

10 Zitat Thomas Schulz und Dr. Wolfgang Stieler: Unter der Haube. Optimierung von CMOS-Transistoren,c´t 5/2000 Seite 260

einwirkende elektrische Feld der Gate-Spannung gesteuert. Oberhalb einer spezifischen Schwellenspannung (...) wird der Kanal leitend und der Transistor schaltet durch.[11]“ Doch bei dieser Beschreibung handelt es sich nur um den, wie erwähnt, prinzipiellen Aufbau beziehungsweise die prinzipielle Funktionsweise. Die Formen von Transistoren, die heute tatsächlich in Prozessoren verwendet werden sind weitaus komplexer. Grund dafür ist die zunehmende Verkleinerung der Strukturen. Problematisch sind vor allem so genannte „heiße“ Elektronen (= besonders energiereich) die den Transistor nachhaltig bis zur Unbrauchbarkeit schädigen können. Um dem entgegen zu wirken, werden schwächer dotierte Zonen eingeführt, der Fachausdruck dazu lautet LDD (Lightly Doped Drain). Diese Gebiete wirken wie Vorwiderstände, die die „heißen“ Elektronen abbremsen. Noch dazu muss man einen zusätzlichen Trick anwenden, um die Schaltgeschwindigkeit zu erhöhen. Eine „ Retrograde Well“ wird eingebaut, die die parasitären Kapazitäten (dadurch wird die Schaltgeschwindigkeit erniedrigt) verringert. Zusätzlich müssen die Leckströme so gut als möglich vermieden werden. Deshalb wird nochmals ein zusätzliches Gebiet auf dem Transistor untergebracht; das so genannte „ pocket“ (auch „halo“ genannt). Der soeben beschriebene tatsächliche Aufbau eines MOSFETs ist also durch die notwendige Verkleinerung der Strukturen bei weitem nicht so einfach wie im Grundkonzept. Um die verschiedenen neuen Teile eines modernen MOSFET noch einmal übersichtlich zu betrachten, sei auf Abb. 7 verwiesen.

11 Zitat Thomas Schulz und Dr.Wolfgang Stieler: Unter der Haube. Optimierung von CMOS-Transistoren, c´t 5/2000 Seite 260/261 (gekürzt und grammatikalisch angepasst)

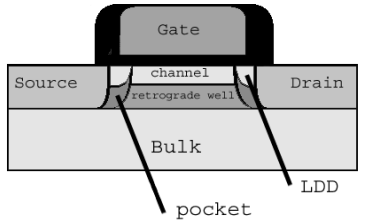


Abb. 7

2.1.2 Die Auswirkungen in der Praxis

Die Verkleinerung der Transistoren und in Folge dessen der Strukturen auf dem Prozessordie ist unbedingt erforderlich, wenn man höhere Geschwindigkeiten erzielen will. Dafür gibt es zwei Gründe: Erstens kann man die Ausmaße des Die nicht beliebig vergrößern, da sie zu schwierig zu kühlen wären; außerdem würde die Produktionsausbeute sinken und dadurch der Preis steigen. Zweitens gilt: je kleiner die Strukturen, desto schneller kann der Transistor schalten. Optimal ist folglich ein kleiner Chip mit einer hoher Packungsdichte. Nachfolgend ist eine Tabelle der aktuellen x86[12] Prozessoren aufgeführt, die die Prozessoren in ihrer Anzahl der Transistoren mit der Chipfläche vergleicht. (der K6-3 läuft hier außer Konkurrenz, da er eigentlich noch aus der 6. Generation von Mikroprozessoren stammt)

12 Siehe Glossar

	Thunderbird	*Athlon*	*Pentium III*	*Pentium III*	*K6-3*
Herstellungsprozess	0,18	0,25	0,18	0,25	0,25
Transistoren	37 Mio	22 Mio	23 Mio	9,5 Mio	21 Mio
Die Größe	120 mm²	184 mm²	106 mm²	140 mm²	118 mm²
Transistoren / mm²	308 333	119 565	216 981	67 857	177 966
L2 onboard	Ja, 256kb	Nein	Ja, 256kb	Nein	Ja, 256kb

Bei genauerer Betrachtung ist eine Sache sehr auffällig. Wenn die L2 Caches Onboard sind, ist die Anzahl der Transistoren pro mm² wesentlich höher. Dies lässt sich erklären, da die Schaltungen für solche Speicherfelder scheinbar einfacher zu designen sind und/oder keine langen Leiterbahnen benötigen. An solchen Vergleichen kann man auch schön die Effektivität eines Prozessordesigns erkennen, oder genauer gesagt dessen Implementation auf dem Die. Diese Fakten haben aber noch nichts mit der tatsächlichen Geschwindigkeit zu tun. Sie sind aber ein Indikator dafür, wie effektiv die Chipdesigner die Architektur auf dem Chip dann wirklich umsetzen konnten. Auch mehr als deutlich ist der Vorteil der kleineren Strukturen zu erkennen, die helfen, mehr Leistung auf weniger Platz zu konzentrieren.

2.2Logisch

Über die logische Umsetzung des Designs eines Mikroprozessors schweigen die Hersteller aus verständlichen Gründen. Sie ist sehr komplex und erfordert unvorstellbares Know how. Bevor es überhaupt zu einer logischen Implementation auf einem real existierenden Chip kommt, wird das neue Design und ihre Architektur beziehungsweise ihre logische Implementation mit einer Simulationssoftware getestet. Dieser Schritt ist sehr wichtig und sollte nie vernachlässigt werden. Eine neue logische Implementation hat immer tausende von Fehlern (in

Fachkreisen wird von „bugs“ gesprochen) die ausgemerzt werden müssen. Viele der Fehler zeigen sich erst viel später und zum Teil bei sehr seltenen Befehlskombinationen. Besonders heimtückisch sind Fehler, die erst bei höheren Taktfrequenzen auftreten, Intel hatte erst vor kurzem einen solchen, äußerst peinlichen Fall mit ihrem 1,1 Ghz Prozessor, der nicht ausgiebig genug getestet wurde und aus Marketinggründen so schnell als möglich auf den Markt kommen sollte. Bei gewissen Befehlskombinationen stürzte der Prozessor reproduzierbar ab, was zur Folge hatte, dass Intel den 1,1 GHz Pentium III Prozessor wieder vom Markt nahm. Solche Fehler passieren immer wieder und sind auch nicht zu vermeiden. Noch dazu sind Simulationen meist sehr langsam, da sie ja im Idealfall jeden einzelnen Transistor und seine Arbeitsweise nachahmen sollten.

2.3 Die Herstellung von Prozessoren und ihre Grenzen

Die Herstellung von Prozessoren ist noch immer eine der größten technologischen Herausforderungen für einen Prozessorhersteller. Die Produktionsausbeute (der „yield“) muss möglichst hoch sein um günstigere und konkurrenzfähigere Produkte auf den Markt zu bringen und die Rohstoffkosten müssen möglichst gering gehalten werden, um den Gewinn zu maximieren. Das folgende Kapitel soll in groben Zügen einen Einblick auf aktuelle Fertigungstechniken in Chipschmieden geben. Das Grundprodukt aller Prozessoren ist bekanntlich Silizium (wobei längst nicht alle Halbleiterbauelemente aus Silizium hergestellt werden). Abb. 8 zeigt Silizium in seinen verschiedenen Formen vom Rohmaterial bis zum Wafer (Rohform in weißem Plastikkübel).

Abb.8

Das Rohmaterial ist zwar in ausreichender Menge vorhanden, jedoch ist die Qualität dessen noch lange nicht geeignet, Chips zu bauen. Es benötigt diverser chemischer Verfahren, um aus dem Rohsilizium hochreines Silizium herzustellen, das für die Produktion von Prozessoren geeignet ist. Das hochreine Silizium allein reicht aber noch immer nicht für die Chipproduktion. Da die Schaltungen und Transistoren immer kleiner werden, ist es notwendig, ein vollkommen gleichmäßiges Kristallgitter zu erzeugen. Dies ist nicht so einfach, wichtig sind vor allem zwei Verfahren, um einen solchen „Monokristall“ zu erhalten: das erste Verfahren heißt „Zonenziehverfahren“. Bei diesem Verfahren wird ein Impfling, von dem aus der Siliziumstab zu „wachsen“ beginnt, an einem Rohsiliziumstab angelagert. „Eine Hochfrequenzspule schmilzt induktiv das Polysilizium an und fährt an diesem entlang. Vom kleinen Monokristall aus ordnen sich nach und nach alle Siliziumatome in einem regelmäßigen Gitter an.“[13]

Das zweite Verfahren wird „ Tiegelziehverfahren“ genannt. „Beim Tiegelziehverfahren erhitzt sich ein keramischer Schmelztiegel mit

13 Zitat Georg Grohs: Chip, Chip, Hurra! Die Herstellung von Halbleiterbauelementen, c´t 24/2000, Seite 275

Bruchstücken aus Reinstsilizium auf 1415 Kelvin. (...) In der Mitte der Schmelze startet ein kleiner Impfling aus Reinstsilizium die Kristallisation. Das Ende des Impflings ist in eine Hebevorrichtung eingespannt. Mit Abnahme der Temperatur und einem vorsichtigen Heben des Impflings lagert sich monokristallines Silizium an den Impfling an.“[13] Die Abb. 9 zeigt einen solchen Einkristall.

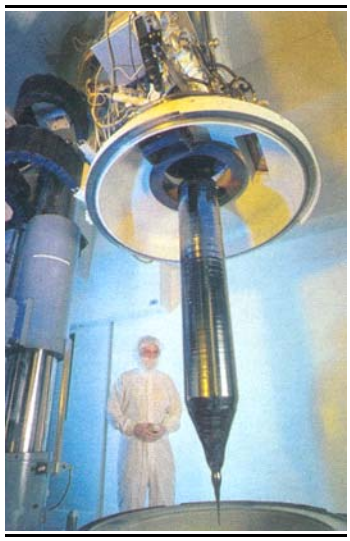


Abb. 9

Bei beiden Verfahren wird der Siliziumstab langsam gedreht, um noch gleichmäßigere Strukturen zu erhalten. Der entstehende Kristall ist aber von seiner äußeren Struktur sehr unregelmäßig. Deshalb muss er in die richtigen Formen gebracht werden. Es ist nicht einfach, einen solchen riesigen Einkristall zu bearbeiten. Er ist sehr spröde und kann bei mechanischer Beanspruchung leicht zerbrechen oder Risse bekommen. Zur Bearbeitung gibt es wieder zwei Möglichkeiten. Der Kristall kann entweder mit einer Innenlochsäge oder mit einer Drahtsäge in einzelne „Wafer“ zerlegt werden. Beide Verfahren haben ihre Vor- und Nachteile, wobei moderne Waferfabriken zunehmend mit Drahtsägen ausgestattet werden. Die durch das Sägen entstandenen Scheiben werden auf ihren

Rändern rundgeschliffen und anschließend an der Oberfläche poliert. Zur Politur von Waferoberflächen gibt es verschiedene Techniken, die auch oft in Kombination angewendet werden. Sie sind: Polieren, Läppen und Feinschleifen. Danach wird die Oberfläche der Wafer nochmals gereinigt, diesmal auf chemischen Wege. Nach der chemischen Reinigung wird der Wafer nochmals chemisch poliert. Zwischen den einzelnen Arbeitsschritten sind natürlich verschiedene Qualitätskontrollen notwendig, um eine einwandfreie Oberfläche zu erhalten. Die fertigen Wafer werden dann in verschiedene Gruppen eingeteilt, je nach ihrem Grad der Abweichungen (die erlaubte Abweichung eines Wafers mit 150 mm Durchmesser beträgt zum Beispiel nur 6µm im polierten Zustand). Jetzt werden die Wafer noch mit verschiedenen Materialien beschichtet, um die vom Chiphersteller gewünschten Eigenschaften zu erhalten.
Zwei Verfahren werden dazu verwendet: Die Dotierung der Halbleiter oder die Epitaxie; bei beiden Prozessen können die Leitfähigkeiten der Elemente stark ins Positive oder ins Negative gewendet werden – je nach Bedürfnis. Dann müssen die Wafer zu den Chipfabriken gebracht werden, was einen unvorstellbaren logistischen Aufwand bedeutet, da die Wafer vollkommen staubfrei transportiert werden müssen und noch dazu keinen nennenswerten Erschütterungen ausgesetzt werden dürfen. Die Wafer werden zwischen den Fabriken berührungslos transportiert, indem sie auf einem Luftpolster schweben. Dann kommen die Wafer zur eigentlichen Chip Herstellung. Sie werden in einem Gestell befestigt und kommen so in die verschiedenen Anlagen zum Auftragen neuer Schichten, Belichten und anschließend Ätzen und Reinigen. Da moderne Prozessoren aus 6 Schichten bestehen, müssen die einzelnen Schritte mindestens 6 mal erfolgen, teilweise sind auch Mehrfachreinigungen

vonnöten.

Im Idealfall kommt dann der Wafer mit den fertigen Chips aus der Produktionsanlage heraus;
Abb. 10 zeigt einen solchen Wafer mit Athlon Prozessoren darauf.

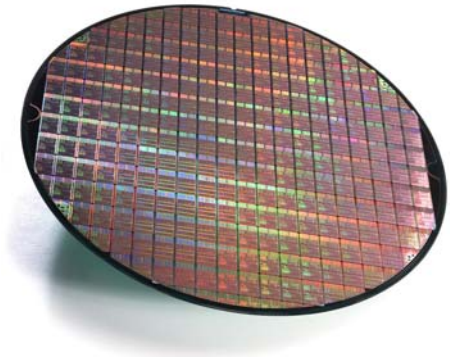


Abb. 10

Wie auf dieser Abbildung deutlich zu erkennen ist, befinden sich auf einem Wafer hunderte von Prozessoren, die erst noch voneinander getrennt werden müssen.[14] Das Wichtigste in einer Chipfabrik (auch „fab“ genannt, wenn sie nur für die eigene Firma erzeugt oder „foundry“ wenn sie auch oder ausschließlich die Chips von Fremdherstellern herstellt) ist die optimale Auslastung der Maschinen, was eine sehr komplexe Regelelektronik benötigt. Teilweise werden die Wafer vollautomatisch zu den einzelnen Produktionsabschnitten gebracht, manchmal müssen aber auch Menschen noch Hand anlegen, meistens aber nur noch zu Wartungszwecken. Danach werden sie in ihre Gehäuse montiert und elektrisch leitend mit den Pins verbunden. Nun werden die

14 Vergleiche Georg Grohs: Chip, Chip, Hurra! Die Herstellung von Halbleiterbauelementen, c´t 24/2000, Seite274f

Prozessoren nach ihrer maximalen Geschwindigkeit getestet. Das bedeutet, die Prozessoren werden eine gewisse Zeit auf eine Art Prüfstand gebracht, wo bei gewissen Geschwindigkeiten das Verhalten des Prozessors bestimmt wird (stürzt der Prozessor bei zum Beispiel 900MHz ab oder wird er zu heiß, wird er als 850MHz Prozessor nochmal getestet). Grundsätzlich ist nur folgender Zusammenhang zu erkennen: Die Prozessoren, die aus der Mitte des Wafers geschnitten wurden, erreichen die höchsten Taktfrequenzen, je weiter außerhalb des Zentrums des Wafers die Prozessoren bei der Belichtung waren, desto langsamer werden sie. Oft planen die Hersteller auch eine gewisse Toleranzgrenze ein oder stufen die Prozessoren absichtlich „falsch“ ein, damit sie den Markt befriedigen können. Dies gibt auch Potenzial für Fälscher und Overclocker, die diese versteckten Leistungspotentiale zu ihren Gunsten ausnutzen. Natürlich kann man keine Architektur endlos ausreizen, irgendwann ist ein Punkt erreicht, an dem eine weitere Verbesserung des Produktionsprozesses nicht mehr möglich ist, obwohl auch in einer einzigen Prozessorgeneration durchaus gewaltige Leistungsunterschiede zu erkennen sind. Da dies nur sinnvoll bei älteren Prozessorgenerationen nachzuweisen ist, sei zum Beispiel die AMD K6 Familie genannt. Ihren Ursprung hatte diese Prozessorgeneration (die sechste von AMD, deshalb K6) im Nexgen Nx586. Die Firma Nexgen wurde von AMD aufgekauft und ihre Technologie im K6 integriert. Anfangs taktete er mit 166 MHz und einer Busfrequenz von 66 MHz. Zum Ende seiner Entwicklung kamen eine SIMD Erweiterung und 256KB onboard Cache (nur K6-3) hinzu, und das bei 100 MHz FSB[15] und bis zu 550 MHz (nur K6-2). Dies wurde vor allem durch die Verbesserung der Herstellungstechnologie ermöglicht, begleitet von einer Verkleinerung der Strukturbreite.

15 Siehe Glossar

Ähnliche Ansätze sind auch schon bei AMD jüngster Prozessorgeneration zu beobachten, die von anfangs 0,25µm auf nun 0.18µm umgestiegen ist. Aber es ergeben sich auch zunehmend Probleme. Es wird immer schwieriger die extreme Hitze, die ein moderner Chip erzeugt, wirkungsvoll abzuführen. Deshalb ergeben sich völlig neuer Ansätze im Chipbereich, wie der Transmeta Crusoe, der ein besonders gutes Verhältnis von Leistung und Verbrauch bietet – dies geschieht aber nicht mehr auf der fertigungstechnologischen Ebene sondern auf der Architekturebene. Ein anderer Ansatz ist die Verwendung neuer Materialien. Schön langsam werden immer mehr Prozessoren in Kupfertechnologie gefertigt. Auch der Athlon Modell 4 wird in Kupfer gefertigt. Der Vorteil von Kupfer ist der geringere spezifische Widerstand gegenüber Aluminium. Der Grund warum man Kupfer nicht schon viel früher verwendet hat liegt darin, dass es Silizium bis zur Unverwendbarkeit verunreinigt. Deshalb müssen spezielle Verfahren zur Anwendung kommen um das Silizium vom Kupfer zu isolieren. Erst vor kurzem ist es den ersten Herstellern gelungen, diese Verfahren auch für die Massenproduktion einzusetzen. Probleme bereiteten vor der Kupfertechnologie die immer höheren MHz zahlen“ Bei diesen hohen Frequenzen sind mittlerweile die Signallaufzeiten in den Leiterbahnen der limitierende Faktor - sie übersteigen die Schaltzeiten der Transistoren bei weitem, so dass bei einer weiteren Erhöhung der Taktfrequenz die Signale nicht mehr zum richtigen Zeitpunkt am Ende der Leiterbahn ankommen.“[16] Zum Vergleich der verschiedenen Technologien sind die Abbildungen Abb.11 und 12 da, wobei die Abb. 11 den herkömmlichen Aluminiumprozess und die Abb.12 den neuartigen Kupferprozess in einer mikroskopischen

16 Zitat Christian Ehmer: PC Professionell 12/1999 S268ff

Aufnahme zeigt.

Abb. 11

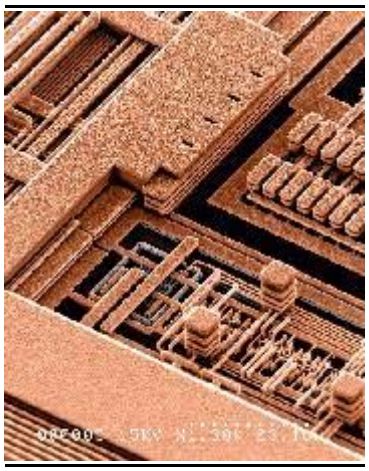


Abb.12

Sehr deutlich ist auf den beiden Fotos auch noch zu erkennen, dass die modernen Prozessoren aus mehreren Lagen Metall bestehen. In Abb. 12 ist die mikroskopische Aufnahme eines PPC, der über sechs solcher Schichten verfügt. Das heißt übrigens nicht, dass der Chip dreidimensional aufgebaut ist, denn es gibt nach wie vor nur eine Schicht mit Transistoren; bei den restlichen 5 Schichten handelt es sich um verschiedene Leiterbahnen, die die einzelnen Transsistoren miteinander

verbinden. Es gibt noch eine weitere Technologie, die in näherer Zukunft für die Herstellung von Prozessoren verwendet werden wird: SOI.
SOI ist eine Abkürzung und bedeutet Silicon On Insulator. An dieser Technik wurde schon seit rund 30 Jahren geforscht, aber bisher ohne Ergebnis. IBM ist es nun erstmals gelungen, SOI auf einem Prozessor funktionsfähig zu implemendieren. Durch SOI werden Effekte vermindert, die durch interne parasitäre Kapazitäten und Transistoren im Silizium entstehen. Solche unerwünschten Effekte erschweren die weitere Verkleinerung von integrierten Schaltungen und infolgedessen verhindern sie eine weitere Geschwindigkeitserhöhung.
Die Abb. 13 zeigt einen Chip, der mit SOI gefertigt ist und auf dem die einzelnen Schichten gekennzeichnet sind.

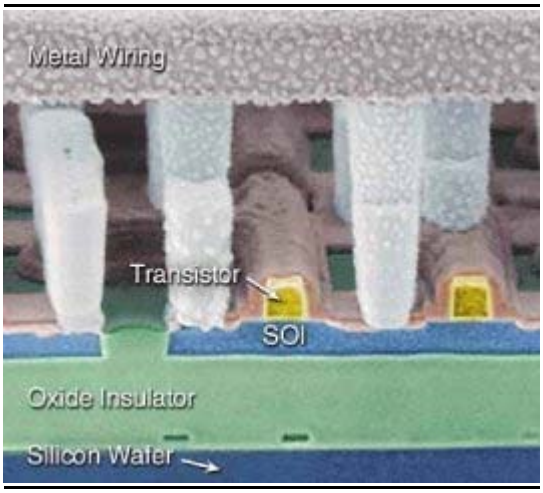


Abb.13

Es gibt auch noch viele andere Möglichkeiten und Versuche, die Strukturen auf Chips weiter zu verkleinern, doch die meisten von ihnen sind noch weit entfernt von der Serienreife.

2.4 Der Von-Neumann-Rechner

Das Von-Neumann Konzept ist das Grundkonzept, auf dem der Großteil

aller modernen Computer aufbaut. „Die kennzeichnenden Merkmale eines Rechners nach dem Von-Neumann-Prinzip sind:

1. Ein (zentralgesteuerter) Rechner ist aus den drei Grundbestandteilen
• Zentraleinheit (Central Processing Unit, CPU)
• Speicher (Memory)
• Ein-/Ausgabeeinheit (Input/Output Unit) aufgebaut. Hinzu kommen noch Verbindungen zwischen diesen Teileinheiten, die als Busse bezeichnet werden. Die CPU übernimmt innerhalb dieser Dreiteilung die Ausführung von Befehlen und enthält die dafür notwendige Ablaufsteuerung. Im Speicher werden sowohl die Daten als auch die Programme in Form von Bitfolgen abgelegt. Die Ein-/Ausgabeeinheit stellt die Verbindung zur Außenwelt in Form des Austausches von Programmen und Daten her.
2. Die Struktur des Rechners ist unabhängig von dem zu bearbeitenden speziellen Problem. Die Anpassung an die Aufgabenstellung erfolgt durch Speicherung eines eigenständigen Programms für jedes neue Problem im Speicher des Rechners. Dieses Programm enthält die notwendigen Informationen für die Steuerung des Rechners. Dieses Grundkonzept der Anpassung hat zu der Bezeichnung 'programmgesteuerter Universalrechner' (engl. 'stored-program machine') geführt.
3. Der Speicher besteht aus Plätzen fester Wortlänge, die einzeln mit Hilfe einer festen Adresse angesprochen werden können. Innerhalb des Speichers befinden sich sowohl Programmanteile als auch Daten, zwischen denen - als Speicherinhalt - grundsätzlich nicht unterschieden wird.

Diese Beschreibungen des Von-Neumann-Rechners führen zu folgender grafischen Darstellung [Abb. 14]

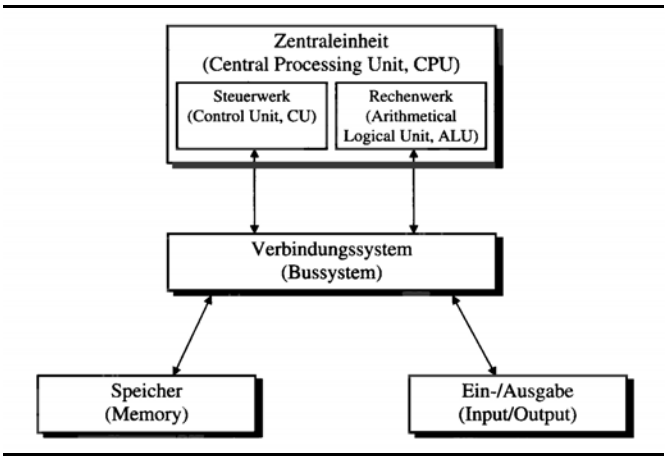


Abb.14

Aus den genannten Grundprinzipien können die wesentlichen Charakteristika des Von-Neumann-Rechners abgeleitet werden:

1. Zu jedem Zeitpunkt führt die CPU exakt einen Befehls aus. Die Steuerung der Bearbeitung liegt im Steuerwerk (Control Unit, CU), das in der Lage sein muss, alle notwendigen Schritte zur vollständigen Behandlung einleiten zu können. Innerhalb eines Befehls kann höchstens ein Datenwert bearbeitet, d.h. neu berechnet werden. Dieses Prinzip wird Single Instruction - Single Data (SISD) genannt.
2. Alle Inhalte von Speicherzellen, im Folgenden Speicherwörter genannt, sind prinzipiell als Daten oder Befehle interpretierbar. Die Daten wiederum können als 'eigentliche' Daten oder als Referenzen auf andere Speicherzellen (Adressen) genutzt werden. Die jeweilige

Verwendung eines Speicherinhalts richtet sich allein nach dem momentanen Kontext des laufenden Programms.
3. Als Konsequenz aus der vorgenannten Eigenschaft können Daten und Befehle nicht gegen ungerechtfertigten Zugriff geschützt werden, da sie gemeinsam ohne Unterscheidungsmöglichkeit im Speicher untergebracht sind. Ein Von-Neumann-Rechner zeigt unter diesen Voraussetzungen einen typischen Verlauf einer Befehlsbearbeitung. Exemplarisch ist in [Abb.15] der Ablauf, insbesondere der Signale auf dem Bussystem, für einen Transferbefehl zwischen Register und Speicher dargestellt.

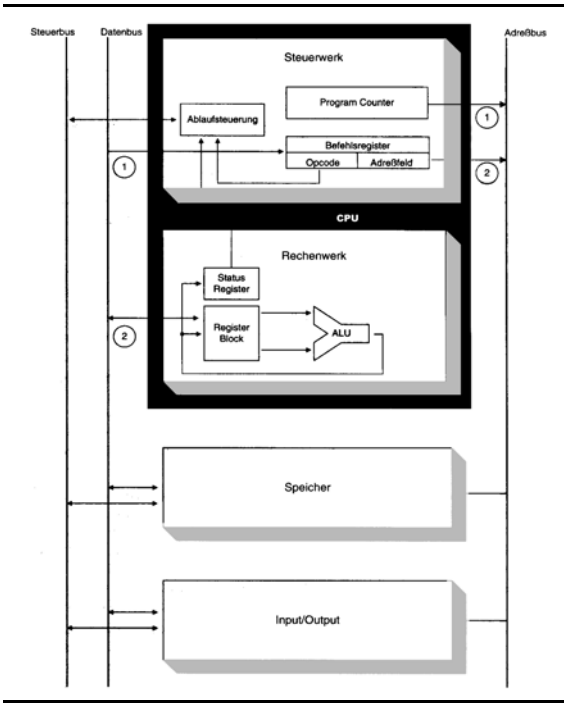


Abb.15

(...)Im Rahmen dieser Befehlsbearbeitung können zwei grobe Phasen identifiziert werden.
In der Phase 1 lädt die CPU, gesteuert durch Abläufe in dem Steuerwerk CU, das Befehlswort aus dem Speicher einschließlich aller benötigten Zusatzinformationen wie Operanden, in diesem Fall aus der

Speicheradresse bestehend. Diese Phase zeigt sich am Bussystem, indem zwei lesende Zugriffe - je einer für das Befehlswort und den Operanden - sequentiell ablaufen. Das Steuerwerk speichert die Inhalte in dafür vorgesehene Befehlsregister und interpretiert sie für die weiteren Aktionen.
In der zweiten Phase wird nun der Inhalt des Registers (...) in die adressierte Speicherzelle kopiert. Hierzu wird der Inhalt aus dem Registerblock des Rechenwerks auf den Datenbus kopiert, während der Adressbus die Information über die gewünschte Speicherzelle trägt. Adress- und Datenbus sind (...) Teilsysteme der Verbindungseinrichtung im Von-Neumann-Rechner. Das hier exemplarisch gezeigte Verfahren zweier Phasen in der Befehlsbearbeitung ist die technische Lösung für das grundsätzliche Problem des Von-Neumann- Rechners, nur eine Verbindungseinrichtung zu zwei verschiedenen Speicherinhalten - Programm und Daten - zu besitzen. Im Allgemeinen gilt hier folgendes Konzept des zeitlichen Multiplex des Bussystems:
1 .In der sogenannten Fetch- und Interpretationsphase wird aufgrund der durch den Befehlszähler angezeigten Adresse der Inhalt einer Speicherzelle geladen und als Befehl interpretiert. Zu dieser Phase zählt im Allgemeinen auch das Laden von Operanden, zumeist von Adressen oder unmittelbaren Daten. Dieser Vorgang wird durch die Interpretation des Befehls gesteuert.
2. In der darauf folgenden Ausführungsphase wird der Befehl nunmehr vollständig interpretiert und ausgeführt. Die Ausführung kann verschiedene Teile der CPU und des Bussystems sowie der angeschlossenen Einheiten in Anspruch nehmen, dies wird durch das Steuerwerk entsprechend gesteuert. Gemeinsam ist diesen Vorgängen, dass alle Speicherzelleninhalte, auch aus dem Ein-/Ausgabesystem, als

Daten interpretiert werden. Dieser zweistufige Ablauf muss für einen Befehl streng sequentiell ablaufen, da eine Abhängigkeit zwischen den verschiedenen Phasen existiert. Spätere Variationen der Von-Neumann-CPU, z.B. die RISC-Architekturen, beinhalten ein Phasenpipelining, das eine scheinbare Parallelität der Aktionen zueinander bewirkt. Dies stellt keinesfalls einen Widerspruch zu dem bisher Gesagten dar, denn die Parallelität im Phasenpipelining bezieht sich auf verschiedene Befehle, während die Bearbeitung eines Befehls weiterhin streng sequentiell bleibt. Der Zeitmultiplex der Busnutzung ist notwendig geworden, da - wie aus dem Beispiel bereits ersichtlich wurde - das Bussystem für den Zugriff auf mehrere Arten von Speicherzelleninhalten genutzt wird. Dies wiederum hat seine Ursache darin, dass der gleiche Speicher fast immer im Mittelpunkt der Operationen steht, unabhängig davon, ob es Programm- oder Dateninhalte sind, auf die zugegriffen werden soll. Die CPU-Speicherkommunikation wird daher die Leistungsfähigkeit des Gesamtsystems entscheidend beeinflussen, was auch als von-Neumann-Flaschenhals (von Neuman bottleneck) bezeichnet wird. - Das von Neumann-Rechnermodell kann als optimal im Sinn von minimal bezeichnet werden. Die in- und externen Ressourcen, die hierbei zur Anwendung kommen, sind nicht weiter minimierbar, ohne eine sehr wesentliche Einschränkung der Funktionalität bis hin zur Sinnlosigkeit des Einsatzes hinzunehmen. “[17]

17 Zitat Christian Siemers: Prozessorbau. eine konstruktive Einführung in das Hardware/Software – Interface Seite 33ff

3 DIE PROZESSORARCHITEKTUR DES AMD ATHLON IM VERGLEICH ZU SEINEN KONKURRENTEN

3.1 Das Design des AMD Athlon

Nun zum zentralen Thema der FBA. In diesem Abschnitt möchte ich die Architektur des AMD Athlon ausführlich präsentieren, ohne aber zu sehr in Details zu gehen, die erstens viel zu schwierig zu verstehen sind und zum Großteil unter Geheimhaltung sind.
Grundsätzlich ist bei Prozessoren eine Einteilung nach ihrem Befehlssatz möglich. Hier gibt es drei wichtige Typen:
RISC, CISC und VLIW.
RISC ist die Abkürzung für Reduced Instruction Set Computing. Dieser Begriff meint nur, dass es sich bei den Befehlen um eine reduzierte Anzahl handelt. Diese Befehle können dann schneller abgearbeitet werden; der Nachteil ist aber, dass manche komplizierte Aufgaben nur mit einer Aneinanderreihung von Befehlen abgearbeitet werden kann.
Hier kommt die CISC Architektur zum Zug:
Die CISC CPU(analog zu RISC: Complex Instruction Set Computing) besitzt wesentlich mehr Befehle als ihre Konkurrenten, und kann dadurch komplexe Befehle schneller ausführen. Doch hier kommt eine oft zitierte Regel zur Anwendung: rund 80 % der durch die Programme aufgerufenen Befehle benötigt lediglich rund 20 % des Befehlsumfangs. So zumindest die Meinung der Experten. Einen völlig anderen Aufbau besitzt der VLIW Prozessor.
Das Besondere an den VLIW Prozessoren ist, dass sie nur etwa zu einem Viertel in Hardware und zu drei Vierteln in Software ausgelegt sind. Ihr erster Vertreter ist der Transmeta Crusoe, auf ihn wird noch im Kapitel 3.2 genauer eingegangen. Der AMD Athlon genauso wie sein kleiner Bruder Duron sind sozusagen Mischlinge aus RISC und CISC Architektur. Extern, also nach außen hin und damit für den Rest der Peripherie und dem Programm, das der Prozessor ausführt, verhält er sich wie ein CISC Prozessor. Intern bearbeitet er die Befehle mit einem RISC Befehlssatz, Bei AMD werden die Befehle MOPS (Macro OperastionS)

beziehungsweise ROPS (RISC OPerationS) genannt; je nachdem wo sich die Befehle gerade „befinden“. Im ersten Moment mag das eigenartig klingen, da die Befehle ja dann kompliziert umgewandelt werden müssen. Doch dieses System hat einen bestechenden Vorteil: die Software, die für ältere Prozessoren geschrieben wurde kann weiter verwendet werden und muss nicht komplett neu kompiliert werden; wenn man dann ältere Software noch nützen will,muss die CISC Umgebung aufwendig emuliert werden, das starke Leistungseinbußen mit sich führen würde. Bei den CISC Befehlen, die der AMD Athlon bearbeiten kann, handelt es sich um den x86 Befehlssatz. x86 bedeutet, dass alle Prozessoren die diesen Befehlssatz verarbeiten können kompatibel zum 8086 sind. So könnte man zum Beispiel auf einem 586 Prozessor theoretisch eine Software ausführen, die für einen 8086 oder 286 geschrieben wurde. Um die Architektur eines Prozessors zu verstehen und eine Übersicht über sie zu bekommen, ist ein so genanntes Blockdiagramm sehr hilfreich. Abb. 16 zeigt das Blockdiagramm des AMD Athlon Modell 1 und Modell 2.

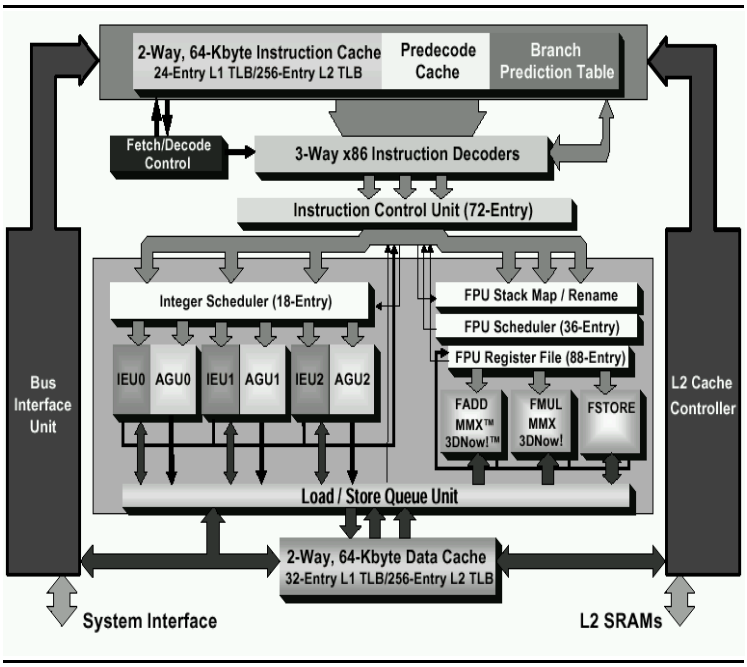


Abb.16

Was an diesem Blockschaltbild zu erkennen ist, sind die wichtigen Verarbeitungseinheiten, von denen es in AMD Athlon 9 Stück gibt. Auf dem Schaltbild sind sie bezeichnet mit: IEU 0, IEU 1, IEU 2, AGU 0, AGU 1, AGU 2, FStore, Fadd/MMX/3Dnow! und Fmul/MMX/3Dnow!. Diese neun Einheiten sind für die Verarbeitung der Befehle mit den dementsprechenden Daten zuständig. „Darüber“ auf der Abbildung befinden sich die Einheiten, die die Befehle an die Ausführungseinheiten verteilen: ICU (Instruction Control Unit), Integer Scheduler und die FPU Stack/Scheduler/Register Einheiten. Nochmal darüber sind die Decoder zu sehen, die den x86 Code in MOPS verwandeln. Die genauen Zusammenhänge zwischen den einzelnen Units sind am besten in der Pipline des Prozessors zu beschreiben. Der AMD Athlon besitzt eine Superscalare Architektur und einen Pipeline Betrieb. Superscalar bedeuted, dass der Prozessor viele verschiedene Ausführungseinheiten besitzt (sie wurden vorher schon erwähnt) und deshalb mehrere Operationen parallel verarbeiten kann. Eine Pipeline kann man sich wie ein Fließband in einer Fabrik vorstellen. Die Abarbeitung eines Befehls wird in viele kleine Unterschritte zerlegt. Jeder der Arbeiter bekommt eine kleine und einfache Aufgabe, die er schnell erfüllen kann, dann rollt das Band weiter. Genauso im Prozessor: Die Abarbeitung eines Befehl wird in viele kleine Unterschritte zerlegt, um diese dann schneller abarbeiten zu können. Weiterhin analog: In der Fabrik ist es möglich, eine einzelne Aufgabe (zum Beispiel die Produktion eines PKW) in unterschiedlich viele Abschnitte zu gliedern. Je kleiner die Abschnitte, desto schneller kann das Förderband fließen. Auch beim Prozessor ist es so: Je mehr Zwischenschritte desto mehr Takt ist möglich. Es gibt nur einen Haken. Was passiert wenn einem der Mitarbeiter ein Fehler unterläuft? Die ganze Produktion muss komplett abgebrochen und neu gestartet werden. Genauso passiert es im Prozessor. Wenn ein Befehl abgebrochen werden muss, da er aufgrund einer falschen „Vermutung“(dazu später mehr) schon bevor der Befehl kam bearbeitet wurde, muss die gesamte Operation abgebrochen und das Ergebnis verworfen werden..
Die Abb. 17 zeigt deutlich die Möglichkeiten den Takt zu erhöhen, wenn

man die einzelnen Stufen zur Befehlsbearbeitung erhöht.

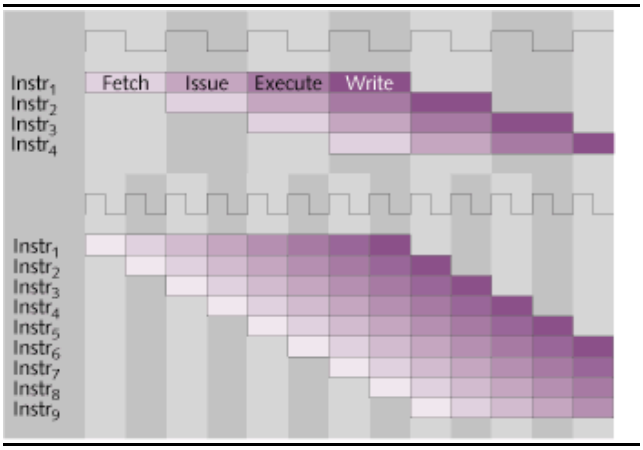


Abb.17

Fetch, Issue usw. sind die einzelnen Befehle einer Pipeline. In diesem Diagramm wurde davon ausgegangen, dass man jede der einzelnen Stationen der oberen Pipeline in zwei kleinere „Unterbefehle“ aufteilt. Dadurch können wesentlich mehr Instruktionen in der selben Zeit abgearbeitet werden. In diesem Fall würde das eine Verdopplung des Taktes zur Folge haben.
Man hat einmal theoretische Überlegungen darüber angestellt, wie lang die optimale Pipeline für einen Prozessor ist. Experten haben ein theoretisches Optimum von etwa acht bis neun einzelner Stufen für klassische Integer Berechnungen herausgefunden. Geringere Piplinelängen würden wie gesagt nur in einem Maße Parallelität der Befehlsabarbeitungen erlauben, wobei die zu fein abgestufte Pipeline viel zu lange Wartezeiten bei einer falschen Vorhersage eines Befehls bedingt. Nun zur Pipeline des AMD Athlon. Die folgende Tabelle zeigt nun erst mal ihre einzelnen Stufen.

Fetch	
Scan	
Align 1	
Align 2	
EDec	
IDec	
Sched (Integer)	*Stack (FP)*
Ex	*Name*
Addr	*WSch*
DC	*Sched*
	FReg
	FX_0
	FX_1
	FX_2
	FX_3

Wie zu sehen ist, verbraucht der Vorgang der Instruktions Decodierung rund die Hälfte der gesamten Pipeline. Mit 10 Stufen im Integer Teil liegt der AMD Athlon damit aber recht nahe am theoretischen Optimum.
Doch nun zur genauen Beschreibung der einzelnen Stufen der Pipeline. Wenn man die einzelnen Stufen der Pipeline betrachtet, kann man auch schön die verschiedenen Einheiten und ihre Funktion beschreiben. Dazu liefert die Abb.18 jedoch vorerst noch einmal ein genaueres Blockbild der AMD Athlon Modell 4. Dieses dient dazu, die Abläufe der Pipeline verständlicher zu machen.

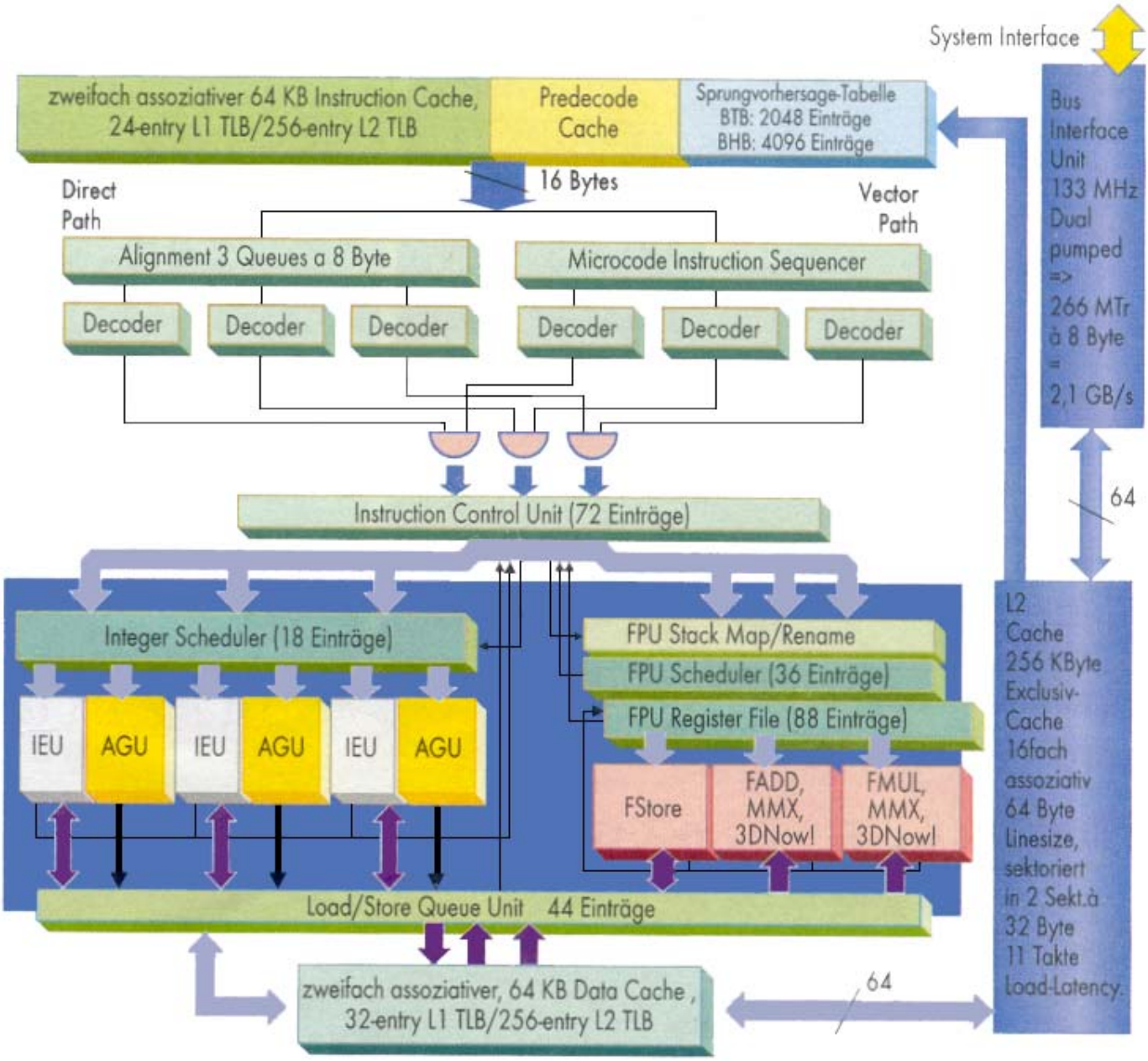

System Interface
zweifach assoziativer 64 KB Instruction Cache, 24-entry L1 TLB/256-entry L2 TLB
Predecode Cache
Sprungvorhersage-Tabelle BTB: 2048 Einträge BHB: 4096 Einträge
Bus Interface Unit 133 MHz Dual pumped => 266 MTr à 8 Byte = 2,1 GB/s
16 Bytes
Direct Path
Vector Path
Alignment 3 Queues a 8 Byte
Microcode Instruction Sequencer
Decoder
Decoder
Decoder
Decoder
Decoder
Decoder
64
Instruction Control Unit (72 Einträge)
Integer Scheduler (18 Einträge)
FPU Stack Map/Rename
FPU Scheduler (36 Einträge)
FPU Register File (88 Einträge)
IEU
AGU
IEU
AGU
IEU
AGU
FStore
FADD, MMX, 3DNow!
FMUL, MMX, 3DNow!
L2 Cache 256 KByte Exclusiv-Cache 16fach assoziativ 64 Byte Linesize, sektoriert in 2 Sekt.à 32 Byte 11 Takte Load-Latency.
Load/Store Queue Unit 44 Einträge
zweifach assoziativer, 64 KB Data Cache , 32-entry L1 TLB/256-entry L2 TLB
64

Abb. 18

Die erste Stufe der Pipeline ist die Fetch Stufe. In diesem Abschnitt werden die Instruktionen aus dem Instruction Cache geholt. Im Instruktionscache können maximal 64 KB an Befehlen zwischengespeichert werden. Maximal können 16 Bytes aus dem Instruktionscache in einem Schritt geholt werden.
Die zweite Stufe der Pipeline heißt Scan; hier werden die x86 Befehle, die aus dem Cache geladen wurden, nach komplexen Instruktionen durchsucht und in das Alignment geschrieben, wenn keine komplexen Instruktionen gefunden wurden. Wenn dies doch der Fall sein sollte, wird der Befehl in den Microcode Instruction Sequencer transferiert. Es gibt folglich zwei Wege für Instruktionen:
Die einfachen Instruktionen werden in den Direct Path geschickt, die komplexen in den Vector Path. (CMPS, CMPSB, CMPSW und CMPSD[18] sind solche komplexe Befehle[19], diese werden aber eher selten gebraucht.). Jetzt kommt auch eine Funktion des AMD Athlon zum Tragen, die vorher schon erwähnt wurde, die Sprungvorhersage (der Mitarbeiter, der einen Fehler macht). Wie schon vorher angedeutet, kann es, während ein Befehl ausgeführt wird zum Abbruch kommen, da sich ein Programm verzweigt (sogenannte „Sprünge“). Wenn dieser Zustand eintrifft, müssen alle Berechnungen verworfen werden; je länger dabei die Pipeline, desto länger braucht es, bis sie wieder gefüllt ist und somit optimal ausgelastet wird. Deshalb verwaltet der AMD Athlon eine Sprungvorhersagetabelle (BPU), die es ihm ermöglicht, solche Sprünge vorherzusagen. Die Sprungvorhersagetabelle besteht aus zwei Teilen; aus der
BHT (Branch History Table, sie speichert das Sprungverhalten an den letzten Sprungadressen) und dem
BTB (Branch Target Buffer, er speichert die Sprungziele ab). Der Beginn dieser Vorhersage ist schon in der 1.Stufe der Pipeline (Fetch). Hier wird

18 Vergleiche:http://developer.intel.com/design/pentium/manuals/24319101.pdf Datum: 20.2.2001

19 Für genauere Informationen und Bedeutung der Befehle:http://developer.intel.com/design/pentium/manuals/24319101.pdf Datum: 20.2.2001

die Adresse des Befehls abgespeichert. Die letztlich endgültige Vorhersage des Befehls wird jedoch in der 2. Stufe der Pipeline gemacht. Die Vorhersagewerte werden in der BHT und dem BTB zwischengespeichert; dafür stehen im BTB 2048 Einträge zur Verfügung, die BHT kann 4096 Einträge verwalten. Für Unterprogramme wird ein kleiner Trick angewendet, der Return-Stack. „In aller Regel werden ja mit CALL aufgerufene Routinen auch mit RET beendet. Der Prozessor merkt sich in einem internen Stack[20] die Rücksprungadresse sowie die Pipeline- und Prefetch-Zustände und muss nicht warten, bis die Rücksprungadresse vom Stack aus dem Cache oder Hauptspeicher eingelesen ist und die Bytes am Sprungziel gelesen und decodiert worden sind.“[21] Dieser Return Stack ist beim Athlon sehr großzügig ausgelegt; er kann maximal 12 Adressen zwischenspeichern.
Die nächsten beiden Stufen (Stufe drei und Stufe vier) heißen Align 1 und Align 2. In diesen Stufen extrahiert ein Verbund von Multiplexern die eintreffenden Instruktionen (sie sind noch immer x86 kompatibel). Multiplexer sind wie folgt definiert:
„Ein n-aus-m-Multiplexer ist eine Funktionseinheit mit n Ausgängen und m Eingängen zum Verarbeiten von Schaltvariablen. Der Wert an einem Ausgang ist abhängig von einer Auswahlfunktion und dem Wert am Eingang der ausgewählten Eingangsvariablen.“[22]
Die Abbildung 19 zeigt ein Schaltbild eines Multiplexers, im Bild handelt es sich um einen 1-aus-4-Multiplexer. Nun zurück zu den Pipelinestufen drei und vier. Die Instruktionen wurden von den Multiplexern angeordnet und zu den eigentlichen x86 Decodern gebracht.

20 Siehe Glossar

21 Zitat Andreas Stiller: Architektur-Wettbewerb. Design-Vergleich: K7-Athlon kontra Pentium III c't 16/1999, Seite 93

22 Zitat Christian Siemers:Prozessorbau, eine konstruktive Einführung in das Hardware/Software – Interface, Seite 15

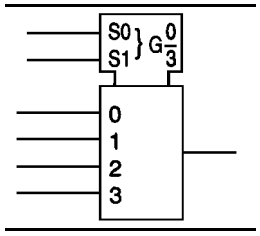


Abb. 19

In Stufe fünf der Pipeline werden die x86 Befehle dann in so genannte MOPS (Marcro OPerationS) transformiert (deshalb wird die Stufe EDec für Early Decode genannt). Die MOPS werden am Ende dieser Phase zur ICU weitergegeben.
Die nächste Phase wird als IDec bezeichnet und ist schon die sechste Stufe der Pipeline des AMD Athlon. Die Stufe sechs läuft in der ICU (Instruction Controll Unit) ab. Sie hat die Möglichkeit, bis zu 72 MOPS zwischenzuspeichern. Die ICU unterscheidet wieder zwei Arten von MOPS. MOPS die Integerberechnungen durchführen und solche, die für die FPU (Floating Point Unit) vorgesehen sind.
Die MOPS für die FPU werden direkt an diese weitergeleitet. Im Unterschied dazu werden die Integer MOPS nicht direkt weitergegeben, sondern ihre Register werden zuerst umbenannt und vom Integer Register werden die Quelloperanden gelesen; an dieser Stelle teilt sich also die Pipeline in FP und Integer Bereich. Bevor die paralell dazu verlaufende FP Pipeline besprochen wird, ist es sinnvoller, sich zuerst nur die Integerpipeline zu Gemüte zu führen.
In Stufe sieben wird der Befehl an den Scheduler der Integer Pipeline weitergegeben. Wenn die ICU einen Befehl zum Integer Scheduler weitergibt, wird er zuerst in eine ROP (Reduced Instruction Set Computig Operation) umgewandelt. ROPs können entweder einen Lade-,einen Speicher-, einen Ladespeicher- oder einen Vorhersagebefehl beinhalten. Ein MOP kann maximal in zwei ROPs umgewandelt werden. So kann zum Beispiel eine Load-Op-Store (=laden, ausführen, abspeichern) x86 Instruktion in eine Load Store ROP und eine ALU

(Aritmetical Logical Unit, führt die Integer Berechnungen durch) ROP zerlegt werden. Der Integer Scheduler kann Operationen auf alle sechs ausführenden Einheiten gleichzeitig (genauer gesagt: in einem Takt) verteilen.
In Stufe acht der Integer Pipeline kommt es dann zur Ausführung des Befehls. Alle Befehle außer Multiplikationen und Divisionen benötigen nur einen Takt. Die Resultate der einzelnen Operationen werden auf einen Result Bus geschickt, von dem aus diese entweder zurück zur ICU kommen oder in der LSU (Load Store Unit) gespeichert werden Die berechnenden Funktionen werden in die ICU zurückgeschrieben, die AGUs (Address-Generation Units) können die Daten direkt in die LSU schicken.
In Stufe neun werden die ausgeführten Befehle, die keine AGU Anweisungen waren, nun von der AGU bearbeitet, deswegen wird diese Pipelinestufe auch als Addr bezeichnet.
In der letzten Stufe (Stufe zehn) der Integer Pipeline werden die fertig berechneten Daten durch die LSU in den L1DCache (L1 Daten Cache) geschrieben. Mit dem Ende dieses Befehls endet auch die Integer Pipeline Nun zur FP Pipeline.

Die FPU Pipeline beginnt an siebenter Stelle der Pipeline. Die FP MOPs werden von der ICU direkt zur FPU weitergegeben.
In Stufe sieben wird der FP Stack umbenannt (deswegen wird dieser Abschnitt Stack/Rename genannt).
In der nächsten Stufe (Stufe acht) werden die Register umbenannt (Rename).
In Stufe neun werden die FP MOPs in einen 36 Einträge fassenden Verteiler transferiert. Diese Einheit, auch FPU Scheduler genannt, ist für die Verteilung der Befehle zu ihren richtigen Ausführungseinheiten da.
In Stufe zehn findet dann die Verteilung statt.
In der nächsten Stufe (Stufe elf) werden die Operanden von den FPU Registern gelesen und die ROPs zu den Ausführungseinheiten gebracht.
In Stufe zwölf beginnt dann die Ausführung der Befehle mit Hilfe der

dementsprechenden Daten aus der LSU. Da FP Operationen eine wesentlich höhere Genauigkeit erfordern, wird der ausführende Teil der FPU Pipeline anders als bei der Integer Pipeline in weitere vier Unterschritte eingeteilt [die Unterschritte werden FX0 (Stufe 12) FX1 (Stufe 13) FX2 (Stufe 14) und FX3 (Stufe 15)genannt]. Somit hat die FPU Pipeline eine Länge von 15 Stufen.[23]

Die FPU besitzt drei vollkommen unabhängige Ausführungseinheiten. Sie werden mit FMul/MMX/3Dnow!, FAdd/MMX/3Dnow!, und Fstore bezeichnet. Jede dieser Einheiten kann in max. vier Pipeline Schritten einen Befehl ausführen (die Einheit FAdd/MMX/3Dnow! kann zum Beispiel in den vier Stufen entweder eine FP Addition, einen MMX Befehl oder einen Befehl aus der SIMD Erweiterung ausführen). Auf 3Dnow! wird in dieser FBA nicht genauer eingegangen; es handelt sich hierbei grob gesagt um eine FP Erweiterung von 45 Befehlen, die alle mit nur einem Befehl aber mehreren Datensätzen arbeiten. Auch die FP Einheiten sind standardisiert. Die FPU des AMD Athlon entspricht im vollen Maße den Standards IEEE 754 und IEEE 854. Diese Standards beschreiben grundlegende Dinge, wie numerische Formate und Ausführungseinheiten. Der AMD Athlon kann mit drei verschiedenen Typen von FP Operationen umgehen.

Im Single-Precision Format erlaubt er bei 32bit Datenweite eine Genauigkeit von sechs bis sieben Stellen.

Im Double-Pecision Format ist eine Genauigkeit von 15 bist 16 Stellen möglich, wobei die Daten 64bit Datenweite besitzen dürfen.

Der Extended-Precision Format Modus erlaubt eine Berechnung mit einer Genauigkeit von 18 bis 20 Stellen, wobei dies bei 80bit Datenweite geschieht.[24]

Wie deutlich zu erkennen ist, besitzt der AMD Athlon eine „mächtige“ Architektur mit sehr viel Parallelität und geringen Ausführungszeiten für Befehle. Doch die „inneren Werte“ allein reichen nicht, um einen Prozessor richtig schnell zu machen.

23 Vergleiche Keith Diefendorff: Microprozessor Report. the insiders´ guide to microprozessor hardware, Volume 12. Number 14, October 26/1998

24 Vergleiche:http://www.amd.com/products/cpg/athlon/pdf/fpu_wp.pdf Datum: 23.2.2001

Hier spielt vor allem die Verbindung mit dem Außensystem eine Rolle (Bus) und das Vorhandensein, die Größe und die Geschwindigkeit von Caches.

3.2Das Bussystem des AMD Athlon

Grundsätzlich sind interne und externe Busse zu unterscheiden. In diesem Kapitel werden jedoch nur die externen Busse kurz erläutert. Technisch ist es nicht machbar, jeder einzelnen Speicherzelle einen eigenen, direkten Zugang zum Prozessor zu verschaffen, da sie viel zu viele Leitungsbahnen erfordern würde. Deshalb gibt es zwischen dem Prozessor und dem Hauptspeicher einen Bus. Er begrenzt zwar einerseits den maximalen Datendurchsatz, ermöglicht aber eine praxistaugliche Lösung des Problemes mit der Verbindung CPU-Speicher. Die effektive Geschwindigkeit eines Busses errechnet sich aus der Anzahl der parallelen Einheiten und der Taktanzahl.
Demnach hat ein Bus der 64bit breit und 100MHz schnell ist eine theoretische Bandbreite von 800MB in der Sekunde (64Bit=8byte; 8 x 100= 800MB) Der Bus des AMD Athlon nutzt zusätzlich eine spezielle Technik: DDR. DDR (Double Data Rate) Busse können in einem Takt 2 mal Daten übertragen, da die Datenübertragung bei steigendem und fallendem Signal getätigt werden. Daraus ergibt sich theoretisch eine doppelte Bandbreite (100MHz DDR entsprechen einem 200MHz Bussystem mit normaler Datenübertragung). Im Fall des AMD Athlon gibt es mittlerweile zwei verschieden schnelle Busse, die aber vom Protokoll her kompatibel sind. Das bringt den Vorteil, dass die Hersteller von Chipsätzen das Know-How von langsameren Implementationen übernehmen können und dadurch wesentlich Zeit sparen. Die zwei verschieden schnellen Busse sind mit 100 und 133 MHz ausgeführt, wobei beide die DDR Technik nutzen. Daraus ergibt sich einen theoretische maximale Geschwindigkeit von 1,6 GB pro Sekunde (100MHz DDR) beziehungsweise 2,1GB pro Sekunde (133MHz DDR). Der Bus hat eine Datenbreite von 64Bit; genau genommen sind es sogar 72Bit, da das Busprotokoll des AMD Athlon noch zusätzlich 8Bit für ein

Fehlerkorrektursystem benutzt. Eigentlich kann man beim AMD Athlon gar nicht von einem Bus sprechen.
Genau genommen kommunizieren Chipsatz und Prozessor mit einen Point-to-Point Protokoll. Dies bedeutet, dass jeder AMD Athlon, der in einem Multiprozessorsystem eingesetzt wird die volle Bandbreite nutzen kann (sofern dies der Chipsatz unterstützt). Dieses Feature wird aber erst bei Multiprozessorsystemen interessant. Das Busprotokoll ist keine AMD Eigenentwicklung, sondern wurde von Digital Equipment lizensiert. Der Vorteil liegt auf der Hand. Digital entwickelt(bzw. entwickelte; Digital wurde mittlerweile von Campaq aufgekauft) schon lange Zeit Rechnersysteme für Hochleistungsaufgaben. Ein vollkommen neues Busprotokoll zu entwerfen und zu bauen erfordert normalerweise viele Jahre, AMD konnte sich diese zeitraubende Aufgabe ersparen.

3.3Die Cachearchitektur des AMD Athlon

Die Caches sind sehr schnelle SRAM Zellen (Static Random Access Memory)die Daten, welche aus dem Hauptspeicher kommen, abspeichern. Diese werden nicht rein zufällig aus dem Hauptspeicher geladen sondern aufgrund von „Vermutungen“. Der AMD Athlon besitzt zwei Abstufungen von Caches.
Der L1 Cache (Level 1 Cache) ist dem Prozessorkern am nächsten außerdem ist er auch am schnellsten.
Der L2 Cache sitzt beim Athlon Modell 4 auf dem Die und wird mit voller Geschwindigkeit angesteuert (also ist der Cache auf dem Prozessor immer gleich schnell wie der Prozessor selbst).
Der L1 Cache ist 128KB groß und zweigeteilt in Instruktionen und Daten (je 64 KB). Um Datenkonflikte zu vermeiden und die Trefferquote des Cache zu erhöhen, hat AMD dem L1 Cache eine zweifache Assoziationseinheit spendiert. Noch dazu ist ein TLB (Translation Lootaside Buffer) vorhanden. Dazu gibt es folgendes vorher zu erklären: der x86 Befehlssatz unterscheidet zwischen logischen und viruellen Adressen. Dies ermöglicht es, größere Mengen an Speicher anzusteuern. Für das Laden und Speichern von Daten aus dem Hauptspeicher müssen daher die viruellen Adressen in physische Adressen umgewandelt

werden. Diese Aufgabe übernimmt die LSU. Durch den TLB kann die Umsetzung wesentlich beschleunigt werden, da er sich die Kombination von virtueller und physischer Adresse in einem insgesamt 56 Einträge fassenden Puffer speichert(24 davon für Daten, 32 für Instruktionen). Für den L2 Cache gibt es einen noch größeren TLB. Er fasst insgesamt 512 Einträge und ist wieder in Instruktions TLB und Daten TLB geteilt (je 256 Einträge). Außerdem sind der L2 Cache 16fach und die beiden L1 Caches je 2fach assoziativ, dieses Feature erhöht nochmals die Effektivität der Caches, indem sie Datenkonflikte vermeidet. Durch die Integration des L2 Cache auf dem Die in AMD Athlon Modell 4 hat AMD auch die maximale Bandbreite erhöht. Laut Angaben von AMD liegt die Steigerung bei 300%.[25]

3.4 Vor- und Nachteile gegenüber anderen Designs

Es gibt viele verschiedene Prozessordesigns, die meisten von ihnen sind aber für den Einsatz in PCs nicht geeignet. Wie schon in Kapitel 1.1 erwähnt, gibt es noch eine große Anzahl von so genannten Embedded Prozessoren (Embedded bedeutet eingebettet). Diese Prozessoren werden in allen Geräten verwendet, die elektronisch gesteuert sind. Dann gibt es noch eine große Anzahl von Prozessoren für Einsatzgebiete, die eine sehr hohe Leistung erfordern und enorm viele einzelne Prozesse gleichzeitig abarbeiten müssen. Der AMD Athlon zielt in der Einprozessor-Konfiguration nicht auf solche Aufgaben ab. Anwendungen, die ein solches Maß an Parallelität erfordern, sind meistens nur bei wissenschaftlichen Simulationen anzutreffen. Die Anwendungen dieser Art werden auf Computersystemen abgearbeitet, die mit der x86 Architektur fast nichts gemeinsam haben. Die Computersysteme bestehen zum Großteil aus vielen CPUs, die alle zueinander parallel arbeiten können. Auch der AMD Athlon hat die Möglichkeit in solche Bereiche vorzudringen, da sein Bussystem sogar, wie vorher erwähnt, aus einem solchen Bereich stammt. Jedoch hat AMD noch keinen Chipsatz auf den Markt gebracht, der die vom Busprotokoll her mögliche Multiprozessorfähigkeit unter Beweis gestellt hat (für den

25 Vergleiche: http://www.amd.com/products/cpg/athlon/pdf/cache_wp.pdf Datum:23.2.2001

AMD 760 ist zwar eine Dualprozessorvariante geplant bis jetzt ist dieser Chip noch nicht in Stückzahlen verfügbar).Aus diesem Grund möchte ich beim Vergleich der verschiedenen Architekturen als erstes die der diversen Hochleistungsprozessoren ausklammern. Die vorher erwähnten Embedded Prozessoren können natürlich auch nicht mit einem AMD Athlon konkurrieren, da in ihren Anwendungsgebieten ein so hohes Maß an Geschwindigkeit und Parallelität nicht erforderlich ist. So bleiben im Groben nur x86 kompatible Prozessoren für den Vergleich übrig (die Architektur der Motorola PPC möchte ich auch weglassen, da sie in ähnlicher Form auch in den Prozessoren für Hochleistungsaufgaben verwendet werden). Die aktuellen Vertreter der x86 Prozessoren sind:
AMD Athlon (Thunderbird / Modell 4)
AMD Duron
AMD Mobile Duron (für Notebooks)
Intel Pentium III (Coppermine)
Intel Celeron
Intel Pentium 4
Intel Mobile Pentium III /Celeron (für Notebooks)
Transmeta Crusoe (für Notebooks)
Via Cyrix III
Zuerst zum Vergleich der verschiedenen Mobilprozessoren. Der Athlon selbst kommt für den Gebrauch in Notebooks nicht in Frage, da er viel zu viel Strom verbraucht. Erst vor kurzem hat aber AMD vom Duron (Athlon mit 64KB L2 Cache) eine Mobilversion herausgebracht, die eine geringere Kernspannung benötigt und daher auch für Notebooks geeignet ist. Intel ist mit den Mobile Pentium III und Mobile Celeron vertreten, sie besitzen die selbe Architektur wie ihre Desktopvarianten, haben aber noch zusätzliche Features zum Stomsparen (Speed Step). Ein Neuling ist der Transmeta Crusoe. Er ist ein Prozessor mit VLIW Architektur, der ähnlich wie der Athlon, nach außen hin x86 kompatibel ist. Der Transmeta Crusor ist der einzige Prozessor, der seine Befehlsabläufe intern optimieren kann und das in Laufzeit. Transmeta nennt dies „Code Morphing“. „Statt starrer Verdrahtung oder wenig flexiblem Microcode vermag die Code Morphing Software (CMS) deutlich mehr, schließlich

ist sie nicht nur wie übliche Microcodes wenige KByte groß, sondern etwa 2 MByte mächtig. So kann sie zur Laufzeit häufig benutzte Sequenzen erkennen und diese Stufe um Stufe optimieren. Manch komplexe x86-Routine mit beispielsweise zehn Befehlen vermag CMS nach mehreren Optimierungsstufen auf nur vier VLIW-'Moleküle' zu reduzieren."[26] (Funktionsschema siehe Abb. 20)

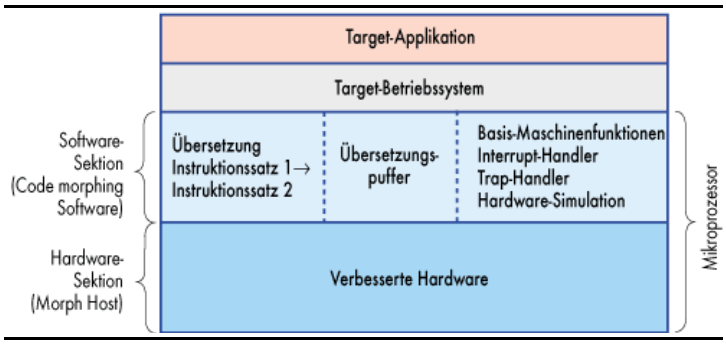


Abb. 20

Noch eine Besonderheit bietet der Crusoe: Auf dem Prozessor ist schon die komplette Northbridge untergebracht, inklusive einem DDR Speicherinterface. Von der Arbeitsweise ist er der einzige Prozessor, der aus dem Schema, das auch der AMD Athlon bietet, deutlich ausbricht. Trotzdem ist auch er genauso ein x86 kompatibler Prozessor, aber aufgrund der flexiblen Architektur kann er sich theoretisch an jede Architektur anpassen (gerüchteweise hat AMD Transmeta beauftragt, einen Emulator auf Basis des Crusoe für den AMD 64Bit Prozessor zu konstruieren, was aber AMD bisher nicht bestätigt hat). Der neue Ansatz von Transmeta ist sehr vielversprechend, hat aber ein großes Problem: Die Systeme mit Crusoe Prozessoren sind einfach zu langsam, um mit anderen Vertretern der Notebook Prozessoren mitzuhalten. Der Vorteil: der Crusoe ist sehr sparsam und wird bei weitem nicht so heiß wie ein Mobile Pentium III von Intel. Die nachfolgenden Bilder sollen dies zeigen. Abb. 21 zeigt den Pentium III in einer Aufnahme einer Wärmekamera beim Abspielen eines DVD Films; Abb.22 dieselbe Tätigkeit auf einem Transmeta Crusoe.

26 Zitat Andreas Stiller: Zu neuen Ufern. Transmetha enthüllt Crusoe-Design c't 3/2000, Seite 32f

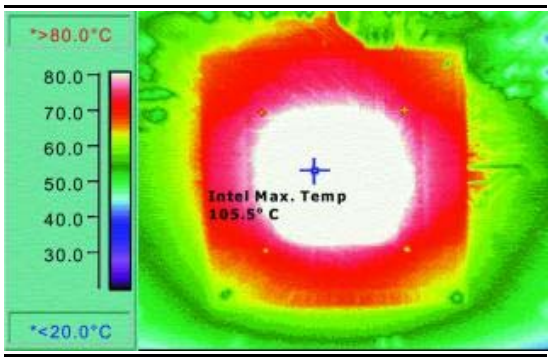


Abb 21

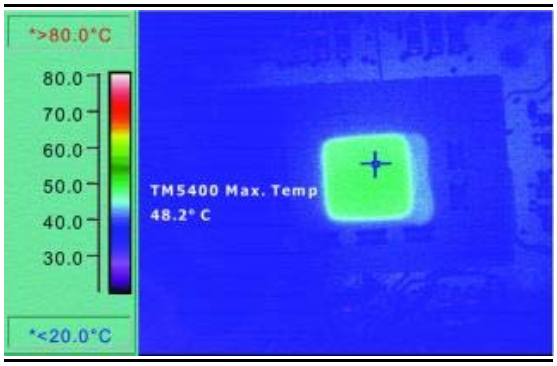


Abb. 22

Doch die ist nur für die Hersteller von Notebooks interessant, dem Anwender ist es vermutlich egal, wie kompliziert die Lüftungstechnik in einem Notebook ist, solange sie funktioniert.

Der Cyrix III bewegt sich an der Grenze von Mobile und Desktop CPU, da er relativ wenig Strom verbraucht, doch auch er bietet sehr wenig Leistung. Der Cyrix III wird wahrscheinlich noch längere Zeit ein Schattendasein fristen, da er wie seine Vorgänger zu wenig Takt und zu wenig FP Leistung erzielte. Aufgrund dieser Rolle wird auf den VIA Cyrix III auch nicht so genau eingegangen, noch dazu bietet sein Design auch keine technischen Raffinessen oder Besonderheiten.

Doch nun zu den Prozessoren für den Desktopbereich. Die wichtigsten Vertreter sind der AMD Athlon und der Intel Pentium III (eventuell auch Pentium 4), eine Preisklasse darunter gibt es nochmals dieselbe Paarung in einer abgespeckten Form (Duron vs. Celeron). Der Duron ist, wie vorher erwähnt, einfach ein AMD Athlon Modell 4 mit einem kleineren L2 Cache; alle anderen Daten sind vollkommen gleich, schwer anzunehmen ist jedoch, dass AMD den

Takt des Duron (derzeit 850 MHz, Stand: 23.2.2001) aus Marketinggründen gering hält. Beim Celeron ist es ähnlich. Auch er ist nur in einer geringeren Taktfrequenz zu erhalten und der L2 Cache ist mit 128 KB nur halb so groß wie die des Pentium III. Einen Unterschied gibt es aber noch: der Bus des Celeron hat nur eine Frequenz von 66 MHz, viel zu langsam für einen modernen Prozessor. Dieses Problem wurde vor kurzem von Intel selbst ausgemerzt: der neue Celeron 800 hat einen FSB von 100MHz, fast soviel wie ein Pentium III, der hat 133MHz Busfrequenz (leider kostete der neue Celeron bei der Präsentation mehr als ein 1Ghz Athlon, für Intel eine peinliche Situation). Nun zum Vergleich AMD Athlon und Intel Pentium III. Zuerst von der Architektur: Der Pentium III besitzt sechs parallele Einheiten, wobei er aber maximal auf fünf Einheiten zeitgleich aus der Reservation Station (entspricht der ICU des AMD Athlon) Befehle verteilen kann; der AMD Athlon hingegen kann alle neun parallele Einheiten gleichzeitig mit Befehlen beliefern. Die Dekodierung der Befehle wird vom AMD Athlon in drei parallelen Befehlsdecodern erledigt, der Pentium III besitzt nur einen. Die folgende Tabelle fasst das eben gesagte kurz zusammen und gibt weitere Auskünfte über die wichtigsten Features der beiden Architekturen.

	AMD Athlon	***Intel Pentium III***
Operationen pro Takteinheit	9	5
Integer Pipelines	3	2
FP Pipelines	3	1
Pipelinelänge (FP/Integer)	(15/10)	(30/17)[27]
FP Operation FADD	4 Takte	3 Takte
FP Operation FMUL	4 Takte	5 Takte
FP Operation FDIV (Single)	16 Takte	17 Takte
FP Operation FDIV(Double)	20 Takte	32 Takte
FP Operation FDIV (Extended)	24 Takte	37 Takte
Vollwertige x86 Decoder:	3	1
Datendurchsatz Bus	2,1GB (FSB133)	1,064GB (FSB133)
SIMD Einheit	3Dnow!	ISSE

27 Diese Angaben stammen nicht von Intel, sondern wurden theoretisch berechnet, da Intel die Pipeline des Pentium III nie veröffentlicht hat.

3.5 Interview mit Jan Gütter, Public Relations Sprecher von AMD

Am 26.1.2001 hatte ich die Gelegenheit, mit dem Public Relations Sprecher von AMD in München zu sprechen. Die wichtigsten Punkte des Gespräches habe ich in folgenden Zeilen niedergeschrieben.

Auch der Athlon hat eine gewisse Entwicklung hinter sich. Vom anfänglichen 0.25 Micron Prozess ausgehend wurde auf 0.18 Micron umgestellt und auch der Cache „wanderte" vom externen Modul direkt auf den Chip. Was ist für die Zukunft geplant?

Der richtige Schritt ist es, den Cache – wie sie es gesagt haben – auf dem Prozessordie und nicht auf dem Board unterzubringen, da die Wege zwischen der kalkulierenden und der speichernden Einheit auf dem Prozessor wesentlich kürzer sind. In Zukunft werden die Bestrebungen immer dahin gehen, den Cache immer so nahe wie möglich auf dem Prozessor unterzubringen; ob das immer möglich sein wird, steht auf einem ganz anderen Blatt. Beim ursprünglichen Athlon war das schlicht und ergreifend nicht möglich, denn ansonsten wäre der Die auch viel zu groß geworden.

AMD hatte schon vor einiger Zeit angekündigt, Prozessoren mit dem Kern des AMD Athlon herauszubringen, die ein bis zwei MB große L2 Caches besitzen, dieses Vorhaben wurde aber verschoben; welche Gründe gibt es dafür?

Es sind weniger technologische Gründe gewesen als vielmehr Marketinggründe; AMD ist dafür bekannt immer in enger

Zusammenarbeit mit seinen Kunden zu sein und auf Kundenseite war die Nachfrage nach derart großen Cache nicht gegeben.

(In Bezug auf den Pentium 4, der eine sehr lange Pipeline besitzt, die eine extrem hohe Taktfrequenz ermöglicht, deshalb aber die schon in den vorhergehenden Kapiteln erwähnten Nachteile aufweist) Wie versucht AMD den Kunden klarzumachen, dass ein Rechner mit 1,2Ghz schneller sein kann als ein PC mit 1,4Ghz? Das sind ja doch sehr komplizierte Zusammenhänge.

All die Leute, die sich für Prozessoren interessieren wissen, dass Taktfrequenz allein nicht alles ist, sondern dass dazu wesentlich mehr gehört. In diesem Zusammenhang hat AMD auch die vergangenen Jahre eine sehr gute Zusammenarbeit mit der Deutschen / Österreichischen / Schweizer Presse geliefert was die Aufklärung über solche Dinge betrifft.

Hat das Busprotokoll von Digital dem AMD Athlon geholfen, sich auf dem Prozessormarkt zu etablieren bzw. wird er das noch tun, z.B. bei den Multiprozessorsystemen?

Am Anfang hat er das sicherlich getan, in der Zwischenzeit hat der AMD Athlon einen sehr guten Ruf was Stabilität und Performance angeht,

und die wenigsten außer die Techniker werden sich für das Busprotokoll interessieren.

Der FSB des AMD Athlon ist derzeit mit 266 MHz getaktet, wird es eine weitere Erhöhung des Busgeschwindigkeit geben, da es das Protokoll ja

ermöglicht?

Wenn es notwendig ist, die Bandbreite des Bus zu erhöhen, wird es sicher geschehen.

Ende diesen, Anfang nächsten Jahres will AMD die ersten 64Bit Prozessoren fertigen, mit diesem Prozessor verlässt AMD erstmals den Weg der Kompatibilität mit Intel, und setzt auf die Eigenentwicklung x86-64 glauben sie, dass das Konzept so überzeugend ist, dass Softwarehersteller für beide Plattformen programmieren werden?

Wir glauben, dass x86-64 ein sehr überzeugendes Konzept ist, nicht nur weil wir das sagen, sondern auch weil das unsere Kunden sagen, niemand ist daran interessiert, die Hard- und Software Infrastruktur die man sich in seinem Konzern aufgebaut hat, über Nacht in den Müllhaufen zu werfen. Deshalb verfolgt AMD einen völlig anderen Ansatz und sagt: wir wollen einen möglichst reibungslosen Übergang zu 64bit, deshalb ist auch im Prozessor 32bit implementiert und diese Prozessoren werden in 32bit nochmals schneller sein als die dann erhältlichen Athlon Prozessoren. (Palomino)Auch von diesem Prozessor wird es verschiedene Versionen geben, die vom Entry Level bist zum High End Bereich alles abdecken.

Werden Die Consumer-Prozessoren dann nach den Serverprozessoren erscheinen?

Es gibt zwei Codenamen: Clawhammer und Sledge Hammer.
Der Clawhammer zielt auf den Consumer Bereich ab, und

der Sledge Hammer ist ein Serverprozessor. Sie werden zumindest zeitgleich erscheinen, wenn nicht der Clawhammer früher eingeführt wird.

Zur Architektur des Sledgehammer. Bei seiner SIMD Erweiterung wurde ja die 3Dnow! Einheit zugunsten der SSE 2 von Intel geopfert; ist das ein Zugeständnis an die Marktmacht von Intel?

Eines vorweg: es gab viele Meldungen dass 3Dnow! nicht in der Hammerfamilie integriert werden soll, das stimmt nicht. Es ist nur SSE 2 folgerichtig hinzugekommen. SSE 2 ist eine durchaus überzeugende Multimediaerweiterung, die von der Industrie gut angenommen wird, deshalb hat sich AMD entschieden diese Mutlimediaeinheit zu unterstützen.

4 ANHANG

4.1 Der Grund dieser Arbeit

Ich habe mich schon lange Zeit mit PC Prozessoren beschäftigt. Mein erstes Magazin, das ich mir von meinem eigenen Geld gekauft habe, war ein Computermagazin, in dem ein großer Vergleichtest von Prozessoren auf der Titelseite angekündigt war. Ich kaufte mir bald darauf alle Magazine, in denen etwas über dieses Thema stand. Nach und nach interessierte mich nicht nur die Leistung des Prozessors sondern auch das, was dahinter steckt, die Architektur eines Prozessors. Mit zunehmendem Interesse an diesem Tema wuchs auch mein Wissen ständig an, und als es zu entscheiden galt, eine FBA über ein Thema zu schreiben, das mich persönlich interessierte und geeignet für eine solche Arbeit war, entschloss ich mich, über die Prozessorarchitektur des AMD Athlon zu berichten. Warum ausgerechnet der AMD Athlon?
Schon zu Beginn meines Interesses galt die Firma AMD als Hersteller von äußerst preisgünstigen Prozessoren, die aber leider einen schlechten Ruf hatten, da sie oft mit Instabilitäten zu kämpfen hatten. Ich entschloss mich aber trotzdem, auf ein AMD System umzusteigen, als ich von einer Ferialarbeit bei Sony das dementsprechende Geld dafür hatte. Ich kaufte mir einen K6-3 450 und konnte mich über Geschwindigkeitsprobleme oder Stabilitätsprobleme nicht beschweren. Die logische Konsequenz daraus war dann vor kurzem der Kauf des zweiten AMD Systems. Diesmal war es ein AMD Athlon mit 1 Ghz und DDR Speicher, der mich im Nachhinein nur noch bekräftigt hat, an der FBA weiterzuschreiben.

4.2Glossar

3Dnow!

SIMD Einheit, die speziell für 3D Anwendungen gedacht ist. Mit dieser Entwicklung kam AMD erstmals Intel zuvor, Intel entwickelte ihre SIMD Einheit erst bedeutend später, sie hieß ISSE.

Embedded

Ein „eingebetteter“ Prozessor, gemeint sind Prozessoren für Handys (DSPs) und andere Anwendungen, die nicht über ein derart breites Leistungsspektrum wie der AMD Athlon verfügen.

FSB

Bedeutet Front Side Bus und ist der Verbindungsbus zwischen CPU und Northbridge. Seine Geschwindigkeit ist ein wesentlicher Bestandteil eines guten Prozessordesigns. Er muss einerseits schnell genug sein um dem Prozessor Daten liefern zu können und andererseits darf er nicht zu kompliziert sein, damit das Boarddesign nicht zu umständlich wird.

Integer

Überwiegend bei Office Anwendungen vorkommende Art von Zahl, die eine feste Kommastelle besitzt, im Gegensatz zur Fließkommazahl, die eine „fließende“ Kommastelle besitzt.

ISSE

Abkürzung für: Internet Streaming Single Instruction Multiple Data Extention, eine von Intel entwickelte Einheit, die es ermöglicht, mehrere Daten mit nur einem Befehl zu bearbeiten.

On Die

Auf dem Prozessordie integrierte Einheit, z.B L2 oder L3 Caches, Eine Northbridge (beim Transmeta Crusoe); On Die Lösungen haben den Vorteil besonders schnell arbeiten zu können, da die Wege zwischen der On Die Einheit und dem Prozessor sehr kurz sind.

RISC

Reduced Instruction Set Computing vereinfachte Befehlssatz, der eine höhere Verarbeitungsgeschwindigkeit zur Folge hat.

Stack

Ein Stack ist ein Last In First Out Speicher. Man kann sich das wie bei einen Tellerstapel vorstellen: die Teller die als letztes auf den Stapel gelegt wurden, können als erste wieder entnommen werden.

x86

Alle Prozessoren, die mit dem 8086 kompatibel sind.

4.3 Literaturverzeichnis

Folgende Quellen wurden für diese Fachbereichsarbeit verwendet:

Christian Siemers: Prozessorbau. Eine konstruktive Einführung in das Hardware / Software Inferface, Hanser Verlag München Wien, München 1999

Bruce D. Shriver: The anatomy of a high-performance microprozessor: a systems perspective, IEEE Computer Society Press Los Alamitos, Los Alamitos 1998

Zusätzlich:

Die c´t Magazine von 2/99 bis 3/2001

Alle anderen Quellen sind direkt im Text angegeben.

Mit der folgenden Unterschrift bestätige ich, keine weiteren Quellen als die Angegebenen verwendet und keine fremde Hilfe beansprucht zu haben .

Alexander Tabakoff, am 25.2.2001

4.4Begleitprotokoll

6. Oktober 1999	***Im Informatikunterricht halte ich ein Referat über die Geschichte der Prozessoren das zwei Stunden dauert. Nach dem Referat spricht mich Prof. Schinwald erstmals darauf an, eventuell eine FBA in Informatik zu schreiben.***
5. Juni 2000	***Prof. Schinwald erinnert mich an die FBA. Ich willige grundsätzlich ein, ohne mich aber auf ein bestimmtes Thema festzulegen.***
13. September 2000	***Ich hade über die Ferien einige Themen zusammengestellt, die von Prof. Schinwald und mir zum Finden des Titels der Arbeit besprochen wurden***
22. September 2000	***Das Thema der Arbeit steht nach einigen Korrekturen fest und lautet nun: „Die Prozessorarchitektur des AMD Athlon“***
7. Oktober 2000	***In einer Buchhandlung in Salzburg erkundige ich mich über vorhandene Fachliteratur.***
9. Oktober 2000	***Ich bestelle das Buch Prozessorbau von Christian Siemers***
8. November 2000	***Besprechung mit Prof. Schinwald***
18. Dezember 2000	***Bestellung des Buches The anatomy of a high-performance microprozessor über Amazon.com***
12. Jänner 2001	***Telefonische Kontaktaufnahme mit AMD***
22. Jänner 2001	***Fertigstellen der Fragen für das Interview***
26. Jänner 2001	***Fahrt nach München Und Interview mit Jan Gütter***
7. Februar 2001	***Letzte Besprechung mit Prof. Schinwald***
26. Februar 2001	***Abgabe der Arbeit***
24. März 2001	***Vorraussichtliche Rückgabe der begutachteten Arbeit***

4.5Bildnachweis:

Titelblatt http://www.amd.com/logos/imagelibrary.html; 10.2.2001

Abb. 1 http://www.amd.com/logos/imagelibrary.html; 10.2.2001

Abb. 2 http://www.amd.com/logos/imagelibrary.html; 10.2.2001

Abb. 3 c´t 14/2000 Seite 32

Abb. 4 http://www.amd.com/logos/imagelibrary.html 10.2.2001

Abb. 5 PC-Welt 11/97 , Seite 250; c´t 5/1999, Seite 119; c´t 14/2000, Seite 95

Abb. 6 c´t 5/2000, Seite 262

Abb. 7 c´t 5/2000, Seite 265

Abb. 8 http://www.wacker-siltronic.com/3/f3.htm 15.2.2001

Abb. 9 c´t 24/2000, Seite 275

Abb. 10 http://www.amd.com/logos/imagelibrary.html 10.2.2001

Abb. 11 http://www.zdnet.de/technik/artikel/cpu/199912/kupfer_01-wc.html 20.2.2001

Abb. 12 http://www.zdnet.de/technik/ artikel/cpu/199912 /kupfer_06-wc.html 20.2.2001

Abb. 13 c't 17/98, Seite 28

Abb. 14 Christian Siemers: Prozessorbau, Seite .34

Abb. 15 Christian Siemers: Prozessorbau, Seite .35

Abb. 16 AMD processor technical documents MED-12/99-0

Abb. 17 Microprozessor Report Volume 12. Number 14, October 26/1998

Abb. 18 c´t 24/2000, Seite 140

Abb. 19 Christian Siemers: Prozessorbau, Seite 15

Abb.20 c´t 25/1999, Seite.26

Abb. 21 Pc Professionell 5/2000 Seite 231

Abb.22 Pc Professionell 5/2000 Seite 231