Hochleistungskommunikationssystem mit Public domain Hardware

Bei der Realisierung von parallelen und verteilten Rechnersystemen ist die Auswahl des Kommunikationssystems entscheidend. Die GMD-FIRST hat für das Gebiet der konfigurierbaren, massiv-parallelen Rechner ein Hochleistungskommunikationssystem entwickelt, das auch in anderen Bereichen mit hohen Kommunikationsanforderungen einsetzbar ist. Das gesamte System und seinen Entwicklungsunterlagen stellen wir jetzt als public domain Hardware zur Verfügung. Dieses Kommunikationssystem kann sehr einfach bei verschiedenen Sorten digitaler Kommunikation verwendet werden. Es erlaubt, die Zahl der Anschlüsse über einen sehr großen Bereich zu wählen, von 16 als kleinster Wert bis zu mehreren Tausenden [http://www.first.gmd.de/~sergio/public.html].

Das Kommunikationssytem wurde bei uns bereits in meheren Projekten erfolgreich eingesetzt. Ursprünglich wurde es für das MANNA-System entwickelt. MANNA (1992) ist ein sehr flexibles, leistungsfähiges und skalierbares MIMD-Testbett für massiv parallele Applikationen. Es ist ein verteiltes Rechnersystem mit verteiltem Speicher. Seine extrem hohe Flexibilität erlaubt dem Benutzer, viele verschiedene Kommunikations- und Programmiermodelle zu testen, ohne dabei auf Leistung verzichten zu müssen. Seine sehr hohe Kommunikations- und Rechenleistung (2 Giga-Flops) ermöglicht es, komplexere Aufgaben in kurzer Zeit zu berechnen, was besonders wichtig ist, wenn man verschiedene Varianten testen möchte oder sehr rechenintensive Aufgaben hat. Ein MANNA-Rechner besteht aus einem Kommunikationssystem, vielen Knotenrechnern und VME-Anschlüssen. Hardware und System-Software sind sehr modular und flexibel, weswegen es auch einfach ist, das MANNA-System von eins bis 16 und dann bis über 4K Knotenrechner und andere Einheiten, mit vielen verschiedenen Konfigurationen und Varianten, zu skalieren.

Das neue Projekt Power-MANNA [http://www.first.gmd.de/PowerMANNA] (1997) basiert auch auf demselben Kommunikatonssystem wie MANNA und seine Knotenrechner basieren auf einem der schnellsten 64-Bit Prozessoren (PowerPC 620). Ein Knotenrechner besteht aus zwei Prozessoren -- einem Kommunikationsprozessor (KP) und einem Applikationsprozessor (AP) -- die mit einem innovativen Speicher-Interface gekoppelt sind. Dadurch ist es möglich, Berechnungen und Kommunikation zu überlappen und so eine noch höhere Parallelität zu erreichen. Die Knotenrechner implementieren ein Pipelined Speicher-Interface mit einer Übertragungsrate von 640 MBytes/s. Dieses Interface wurde auch in einem eigenen Chip implementiert.

Die Grundbausteine des Kommunikationssystems sind ein 16x16-Crossbar und ein universeller Anschluß (Link-Anschluß) für die Endgeräte. An jeden Anschluß kann ein Knotenrechner, ein I/O-Device oder ein Interface zu einem Standard Bus angeschlossen werden. Die Link-Anschlüsse sind bidirektional und vollduplex. Jeder Anschluß besitzt eine Transferrate von 50 bis 60 MBytes/s für jede Richtung (100 bis 120 MBytes/s vollduplex). Das Netzwerk hat einen integrierten Router, um selbständig die Nachrichten an ihr Ziel zu bringen. Kleine Systeme (mit 16 oder weniger

Jara, Montenegro & Mousa; GMD-FIRST, Rudower Chaussee 5;
D-12489 Berlin-Adlershof
Tel +49/30/6392-1878, email: sergio@first.gmd.de

Anschlüssen) haben eine Routing-Zeit von weniger als einer halben Mikrosekunde. Große Systeme mit mehreren Tausend Knoten haben eine Routing-Zeit von 2 bis 3 Mikrosekunden. Im Netz können viele Verbindungen gleichzeitig existieren, ohne sich gegenseitig zu stören oder Bandbreite zu stehlen.

Um verschiedene Einheiten problemlos miteinander verbinden zu können, wurde ein universeller Verbindungskanal ("LINK") geschaffen und für diesen ein "LINK-Protokoll" definiert. Dieses Protokoll wird nicht nur benutzt, um die verschiedenen Einheiten mit dem Netzwerk zu verbinden, sondern auch, um netzwerkintern die verschiedenen Netzkomponenten zu koppeln. Auf diese Weise ist es möglich, jede Einheit unmittelbar mit jeder anderen zu verbinden (Abbildung 1).

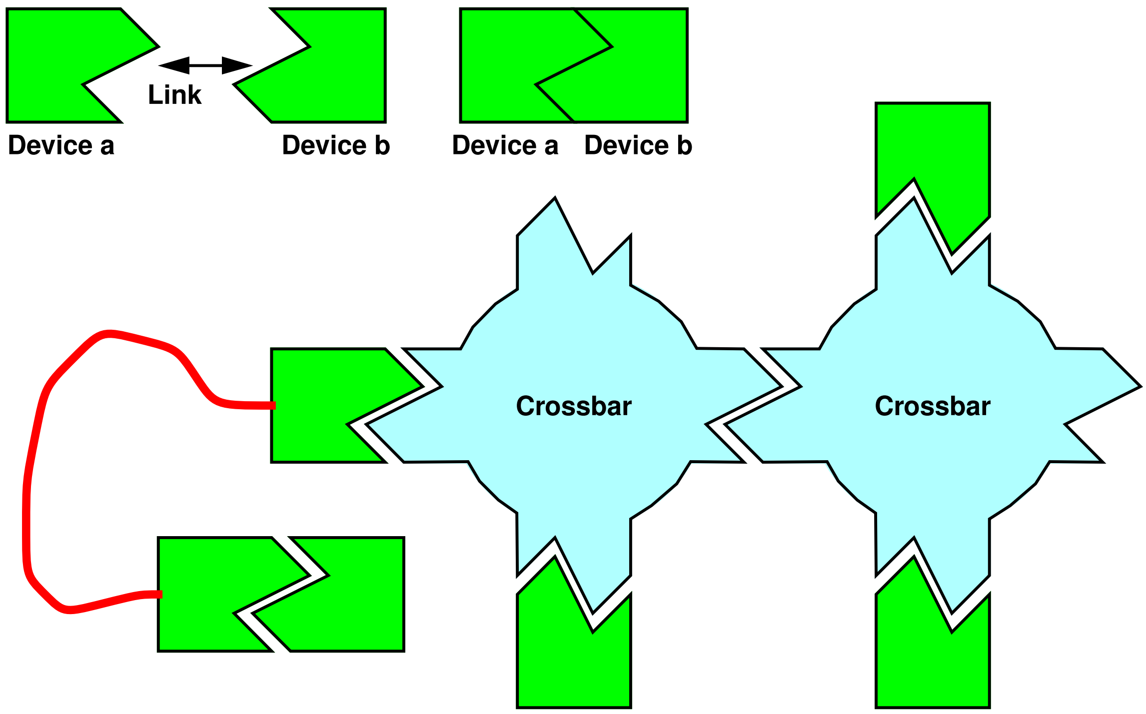

Abbildung 1

Das Verbindungsnetzwerk ist eine Hierarchie von 16x16-Crossbars, die alle auf der Grundlage des LINK-Protokolls arbeiten. Die Topologie des Netzwerkes, die Redundanz der Verbindungen und die Anzahl der Anschlüsse im Netzwerk sind frei konfigurierbar; die Anschlüsse können wahlweise für Rechenknoten oder andere Einheiten benutzt werden. Dazu wird ein softwaregesteuertes Verfahren des in Hardware implementierten Routers benutzt. Das Verbindungsnetzwerk erhält dadurch eine fast beliebige Flexibilität und Skalierbarkeit und kann zur Erprobung beliebiger Kommunikationsstrukturen dienen.

Jara, Montenegro & Mousa; GMD-FIRST, Rudower Chaussee 5;
D-12489 Berlin-Adlershof
Tel +49/30/6392-1878, email: sergio@first.gmd.de

Link-Protokoll

Der LINK ist eine taktsynchrone, byte-parallele, bidirektionale Punkt-zu-Punkt Verbindung, die mit einer Taktfrequenz von 50 MHz arbeitet. Jeder Anschluß kann gleichzeitig eine eingehende und eine ausgehende logische Verbindung von jeweils 50 Mbytes/s (100 Mbytes/s vollduplex) herstellen. Diese Eigenschaft verbessert nicht nur die Transferrate, sondern vereinfacht auch die Verbindungsaufbauprotokolle, indem Verbindungsverklemmungen (*deadlocks*) ausgeschlossen werden. Da sich alle Verbindungen im System an das LINK-Protokoll halten, ist es möglich, jede Einheit mit jeder beliebigen anderen zusammenzuschalten. Damit lassen sich beliebigen Verbindungsstrukturen aufbauen. Durch den LINK sind die Einheiten unabhängig von Netz, Struktur und Größe.

Die Links sind in der Lage, gleichzeitig Daten und Kontrollinformationen in beide Richtungen zu übertragen. Jede Richtung eines LINKs besitzt einen 9-Bit breiten Kanal für die Daten und Kontrollinformationen zwischen Sender und Empfänger und ein *stop*-Signal für eine weiche Datenflußkontrolle in die entgegengesetzte Richtung (Abbildung 2). Auf der Empfängerseite eines LINKs ist ein FIFO-Puffer. Solange genügend Platz im Puffer vorhanden ist, kann der Sender mit voller Geschwindigkeit senden (ein 9-Bit Wort pro Taktzyklus). Wird der freie Platz im Puffer knapp, setzt der Empfänger das *stop*-Signal, was zur Unterbrechung des Sendens führt.

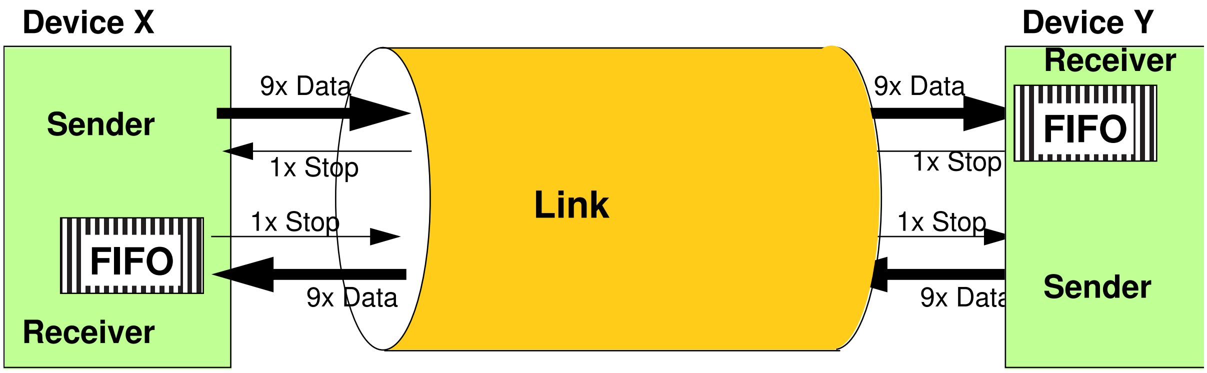

Abbildung 2: Link

Crossbar

Um mehrere LINK-Anschlüsse vernetzen zu können, wurde ein Crossbar mit 16 Eingangs- und 16 Ausgangskanälen entwickelt und mit dem vollständigen LINK-Protokoll als ASIC realisiert. Intern im Crossbar ist eine vollständige Vernetzung der Eingänge und Ausgänge implementiert, so daß jeder Eingang alle Ausgänge direkt erreichen kann. Alle diese Verbindungen sind unabhängig voneinander und können daher gleichzeitig, und ohne sich gegenseitig zu stören, hergestellt werden. Jeweils ein Eingang und ein Ausgang können zusammen verwendet werden, um einen Full Duplex-Anschluß zu realisieren, wie es das LINK-Protokoll erfordert (Abbildung 3).

Die FIFOs in den Eingangskanälen sind 32 Bytes (32x9-Bit) tief und werden für eine “weiche” Datenflußkontrolle benutzt. Mittels eines Routers, der sich in jedem Eingangskanal befindet, kann der Crossbar intern jeden Eingang mit jedem Ausgang verbinden. Es besteht dann eine logische Verbindung zwischen dem Eingangs- und dem Ausgangskanal. Der Verbindungsaufbau benötigt im kollisionsfreien Fall zwischen 0.2 und 0.5 Mikrosekunden. Eine logische Verbindung wird durch ein *route*-Kommando hergestellt. Von da ab können beliebige Kommandos und Datenpakete übertragen werden. Am Ende wird die Verbindung mit einem *close*-Kommando geschlossen.

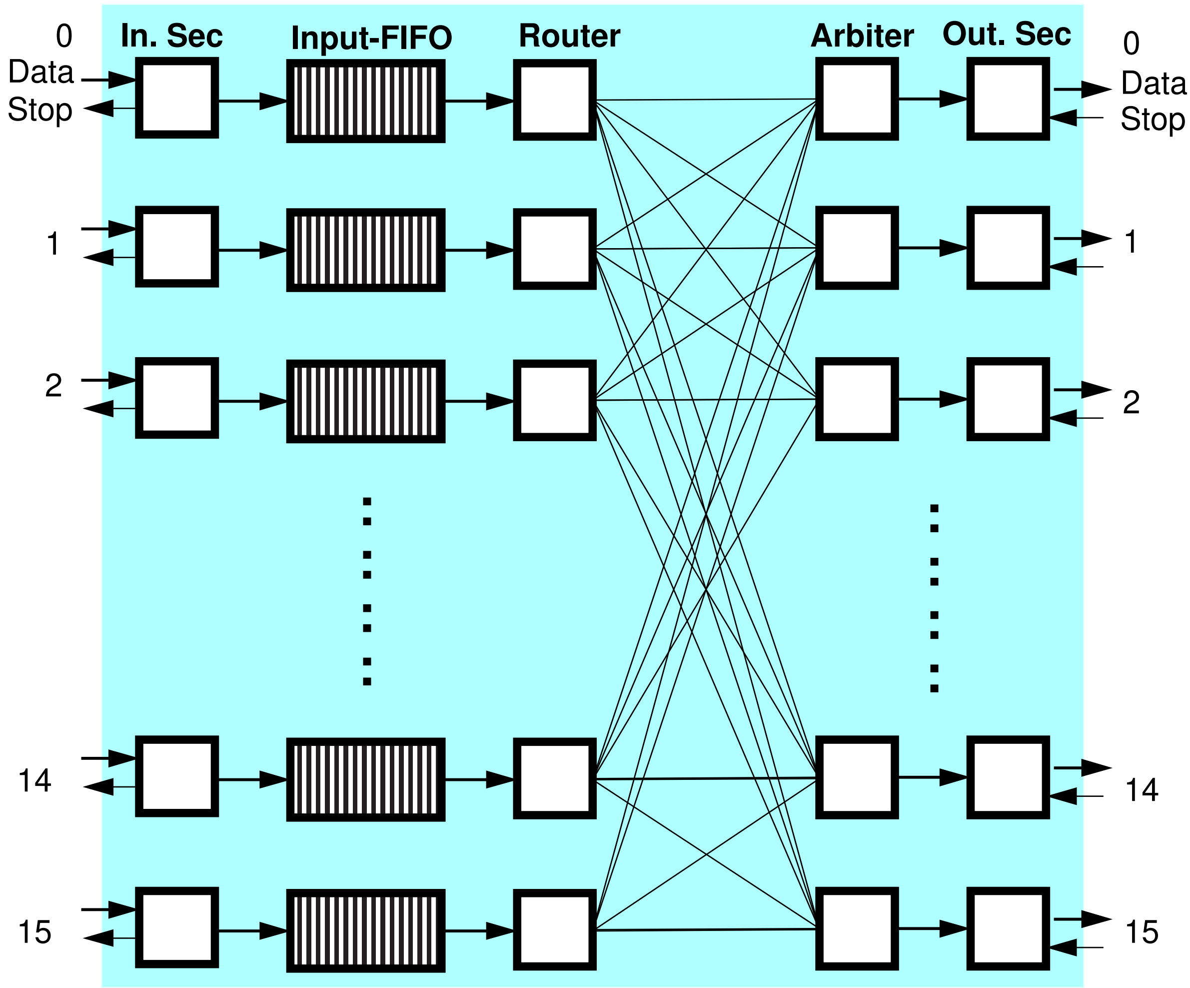

Abbildung 3: 16x16 Crossbar

Asynchrone Strecke

Da das LINK-Protokoll taktsynchron ist und mit einer hohen Frequenz (50 MHz) arbeitet, müssen die Verbindungen physikalisch kurz sein (bis ca. 50 cm). Um Einheiten, die weiter voneinander entfernt sind oder mit verschiedener Frequenz arbeiten, miteinander zu verbinden, wurde die *asynchrone Strecke* entwickelt.

Jara, Montenegro & Mousa; GMD-FIRST, Rudower Chaussee 5;
D-12489 Berlin-Adlershof
Tel +49/30/6392-1878, email: sergio@first.gmd.de

Eingang und Ausgang der asynchrone Strecke arbeiten nach dem LINK-Protokoll. Beide Enden werden durch parallele ECL-Signale miteinander verbunden. Damit kann ein Abstand bis zu 10 m überbrückt werden.

Die asynchrone Strecke implementiert das vollständige LINK-Protokoll. Die Empfänger an den Streckenenden verfügen über hinreichend große FIFO-Puffer, um eine “weiche” Datenflußkontrolle über lange Verbindungen garantieren und gegebenenfalls verschiedene oder verschobene Takt-Frequenzen der beiden Enden kompensieren zu können (Abbildung 4).

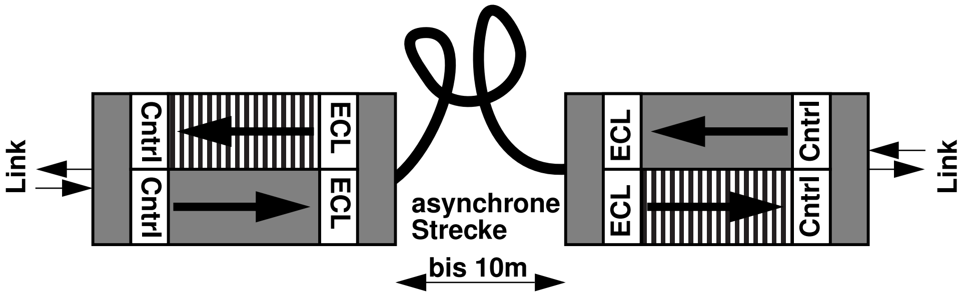

Abbildung 4: Asynchrone Strecke

Link Interface

Um einen einfachen und kompakten Anschluß einer 64-Bit Einheit (z.B. Knotenrechner) zum LINK-Protokoll zu ermöglichen, wurde ein LINK-Interface-Chip entwickelt. Außer der Protokoll-Konvertierung und dem Datentransfer führt das LINK-Interface zur Erhöhung der Zuverlässigkeit der Nachrichten auch eine CRC-Checksummen-Generierung und Prüfung durch.

Das LINK-Interface unterstützt die gesamte Funktionalität des LINK-Protokolls. Es hat einen Sende- und einen Empfangskanal (Abbildung 5). Beide Kanäle sind voneinander unabhängig und können simultan arbeiten. Jede Richtung hat einen FIFO-Puffer, der bis zu 32 64-Bit-Wörter (256 Bytes) fassen kann. Der FIFO wird benutzt, um beide Seiten des LINK-Interface (CPU und Crossbar), die unterschiedliche Datenübertragungsraten besitzen, zu entkoppeln. Der Sende-FIFO kann außerdem kleine Nachrichten (bis zu 256 Bytes) zwischenspeichern, so daß der Sendeprozeß nicht auf den Verbindungsaufbau zum Empfänger warten muß.

Beide Seiten des Interface-Chips sind taktsynchron und können ein Wort pro Takt übertragen: 64 Bit mit einer Übertragungsrate von 400 MBytes/s auf der CPU-Seite und 8 Bit mit einer Übertragungsrate von 50 MBytes/s pro Richtung auf der Link Seite.

Jara, Montenegro & Mousa; GMD-FIRST, Rudower Chaussee 5;
D-12489 Berlin-Adlershof
Tel +49/30/6392-1878, email: sergio@first.gmd.de

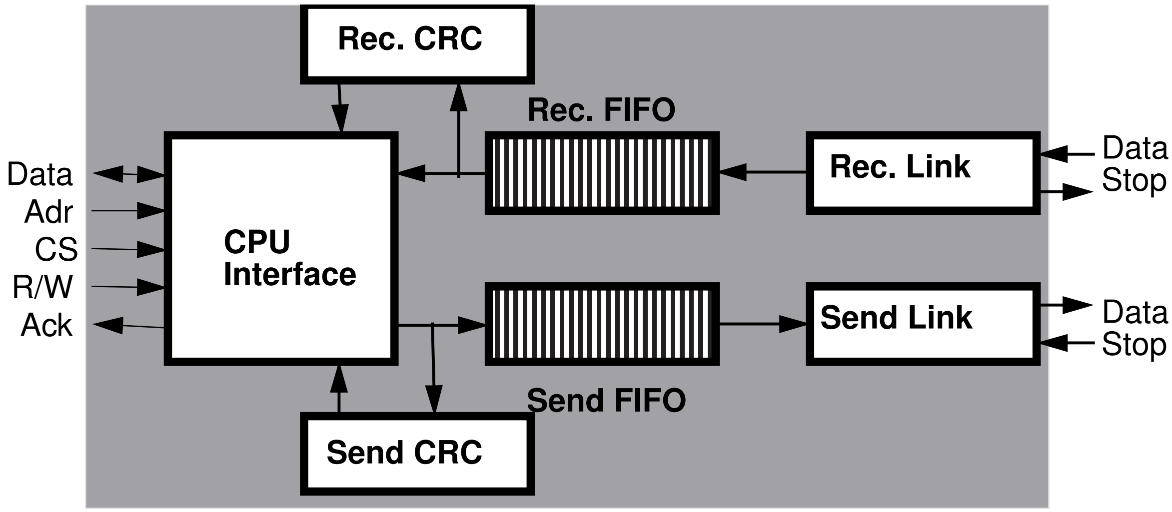

Abbildung 5: Link Interface

Topologie

Die effektive Leistung eines Parallelrechners hängt wesentlich von der Leistungsfähigkeit des Kommunikationssystems ab. Deshalb wurden in der Vergangenheit oft komplexe und teure Verbindungstopologien, wie z.B. der Hypercube, benutzt. Für massiv-parallele Systeme ist eine solche Topologie aber kaum mehr realisierbar und vor allem nicht wirtschaftlich. Deshalb konzentrierten sich unsere Arbeiten auf die Entwicklung neuartiger Topologien, um gleich gute Verbindungseigenschaften wie beim Hypercube mit einfacheren Strukturen zu erreichen.

Um die Komplexität zu senken, wurden so viele Verbindungen wie möglich integriert. Die erste Integrationsstufe ist der 16x16 Ein-Chip-Crossbar, der 265 Verbindungen ersetzt. VLSI-Elemente sind billiger, kleiner, schneller und sicherer als Kabel. Kleine Systeme (“Super-Arbeitsplatzrechner”) mit bis zu 16 Knoten lassen sich auf einfache Weise realisieren, indem man die Knoten über einen solchen Crossbar zu einem Cluster zusammenschaltet. Die zweite Verbindungssstufe kann durch eine backplane mit 2 oder 4 Crossbars und 20 oder 40 Anschlüsse.

Große Systeme werden realisiert, indem mehrere Crossbars zusammengeschaltet werden. Dafür sind beliebige Konfigurationen machbar. Es ist jedoch sinnvoll, die passende Topologie für eine bestimmte Systemgröße zu realisieren. Diese Struktur wird nicht immer gleich sein, da sie für eine gegebene Systemgröße optimiert wird.

Backplane

Wir haben eine Backplane mit zwei Crossbars, 16 synchrone Anschlüsse und 16 asynchrone Strecken entwickelt, und bieten sie auch als Public Domain Hardware an (Siehe Abbildung 6). Dies System kann mit bis zu 70 MHz laufen.

Jara, Montenegro & Mousa; GMD-FIRST, Rudower Chaussee 5;
D-12489 Berlin-Adlershof
Tel +49/30/6392-1878, email: sergio@first.gmd.de

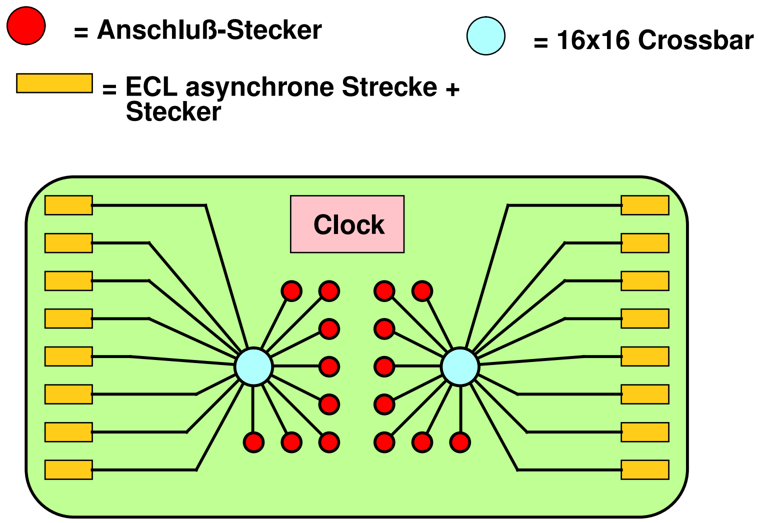

Abbildung 6: Backplane

Public Domain Hardware

Die benötigten Entwicklungsunterlagen sind bei der URL
http://www.first.gmd.de/~sergio/public_domain/
erhältlich. Für die Rechte über Chips und Platinen, sowie Schaltpläne und andere Entwicklungsunterlagen, bitte kontaktieren Sie uns:

Sergio Montenegro	Fihmi Mousa	Elizardo Jara
sergio@first.gmd.de	fihmi@first.gmd.de	jara@first.gmd.de
Tel 030/6392-1878	Tel 030/6392-1838	tel 030/6392-1884