CRAY T3E Aufträge in Deutschland

- **Flüssiggekühlte Systeme**

  LC512 Forschungszentrum Jülich

  LC416 MPG - Rechenzentrum Garching

  LC128 Konrad-Zuse-Zentrum für Informationstechnik Berlin

  LC512 Rechenzentrum Universität Stuttgart

- **Luftgekühlte Systeme**

  AC MPI - Göttingen, MPI - Berlin, MPI - Mainz, MPI - Stuttgart

  AC16 Technische Universität Berlin (Fakultät für Informatik)

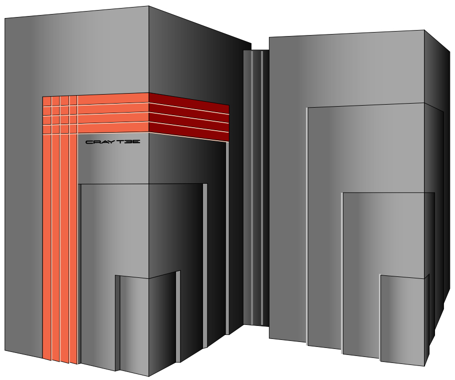
  AC16 RRZN - Universität Hannover

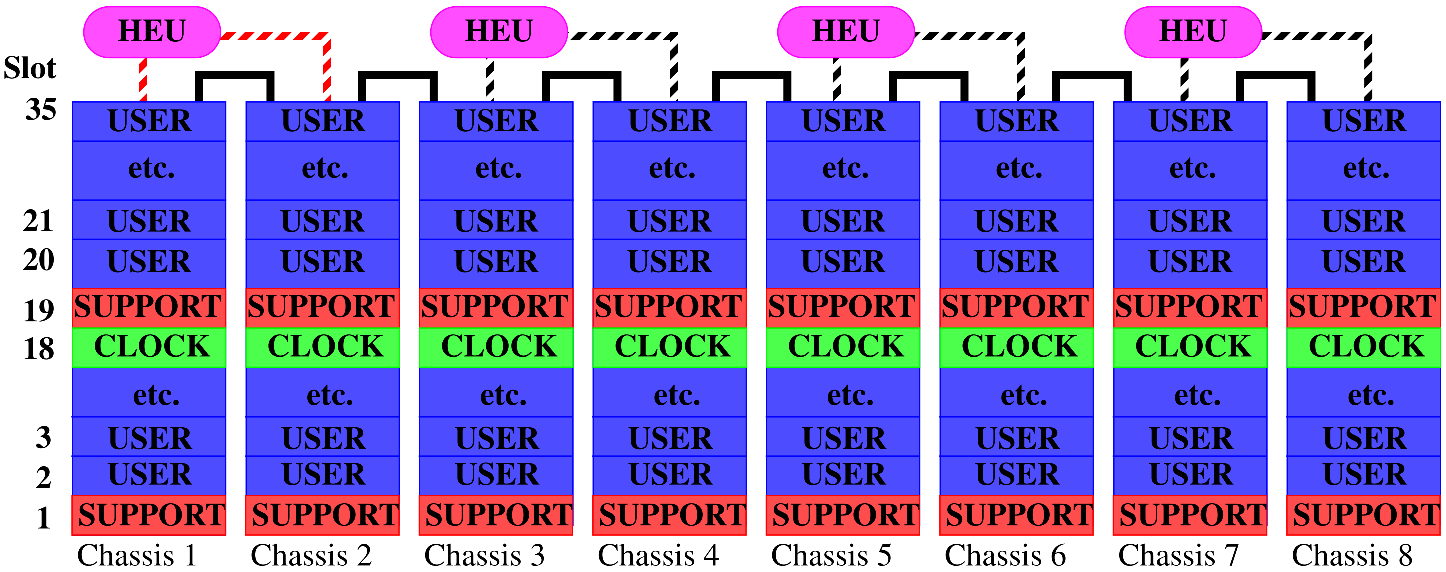
  AC56 Alfred Wegener Institut für Polarforschung Bremerhaven

  AC32 Universität Rostock

Flüssiggekühlte CRAY T3E Konfigurationen

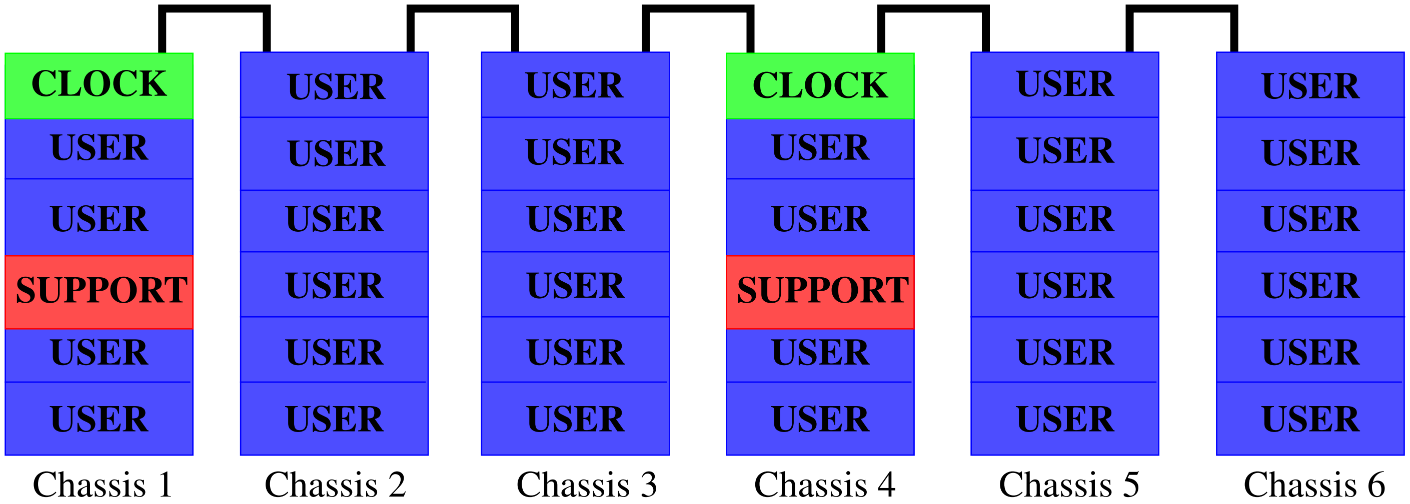
- **2 Printed Circuit Boards (je 4 PEs und 1 I/O-Kontrolleinheit) pro Modul**
- **Aufrüstbarkeit in Schritten von einem Modul (8 PEs)**
- **Chassis-Größe (in cm) ca. B x T x H = 120 x 180 x 180**
- **64 bis 256 Anwender-PEs, zusätzlich 8 oder 16 System-PEs pro Chassis**
- **1 GigaRing Kanal pro 16 User PEs; 16 Bulk Heads pro Chassis**
- **Konfigurationen von 64 bis zu 2048 PEs**



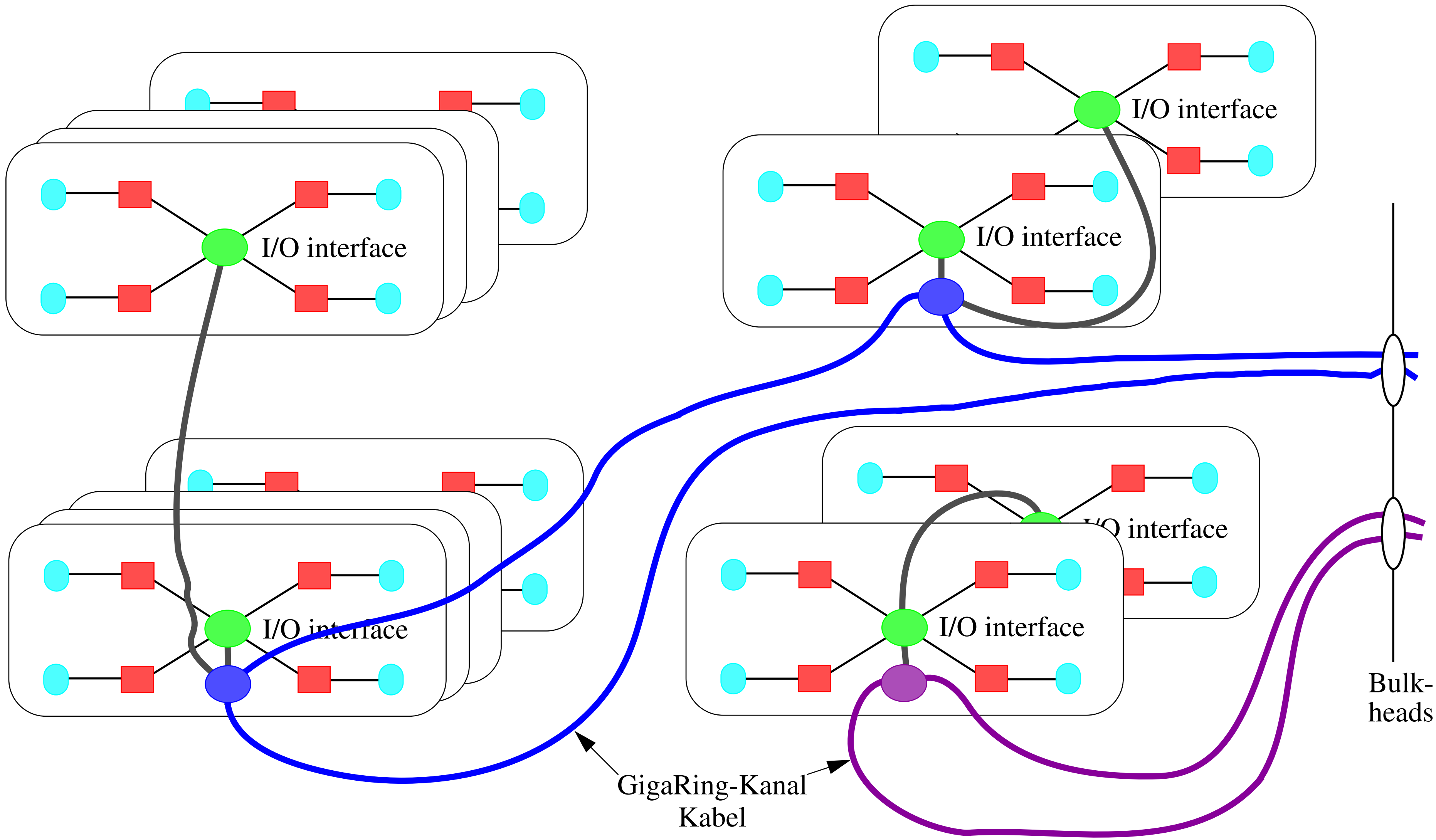


Luftgekühlte CRAY T3E Konfigurationen

- **1 Printed Circuit Board (4 PEs und 1 I/O-Kontrolleinheit) pro Modul**
- **Aufrüstbarkeit in Schritten von einem Modul (4 PEs)**
- **Chassis-Größe (in cm) ca. B x T x H = 60 x 150 x 175**
- **20 (plus 1 Clock-Modul) oder 24 PEs in einem Chassis**
- **1 GigaRing Kanal pro 8 User PEs; 3 Bulk Heads pro Chassis**
- **10 bis 128 Anwender-PEs, zusätzlich 2, 4 oder 8 System-PEs**

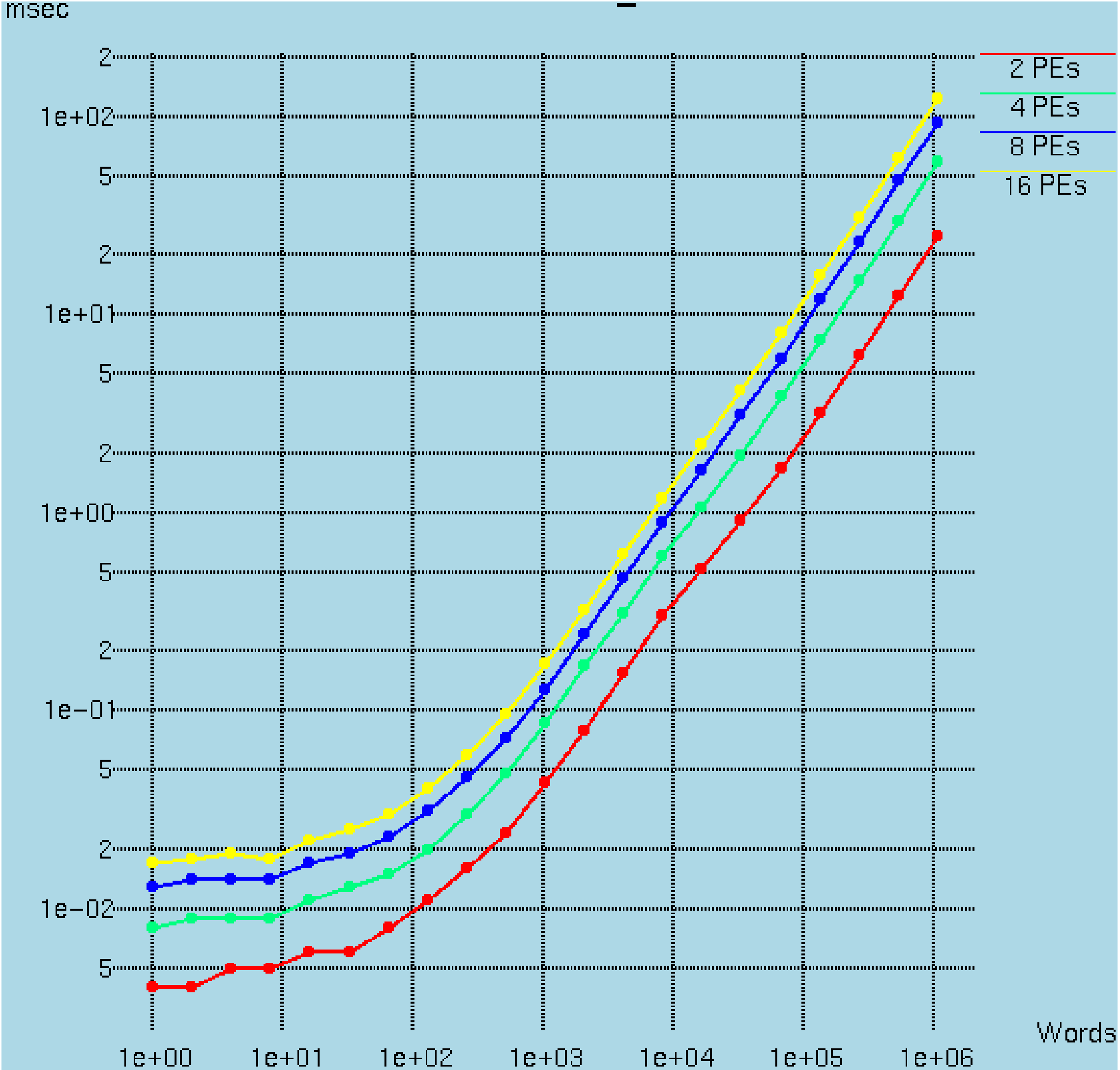


CRAY T3E - I/O-Interface zu GigaRing Kanal



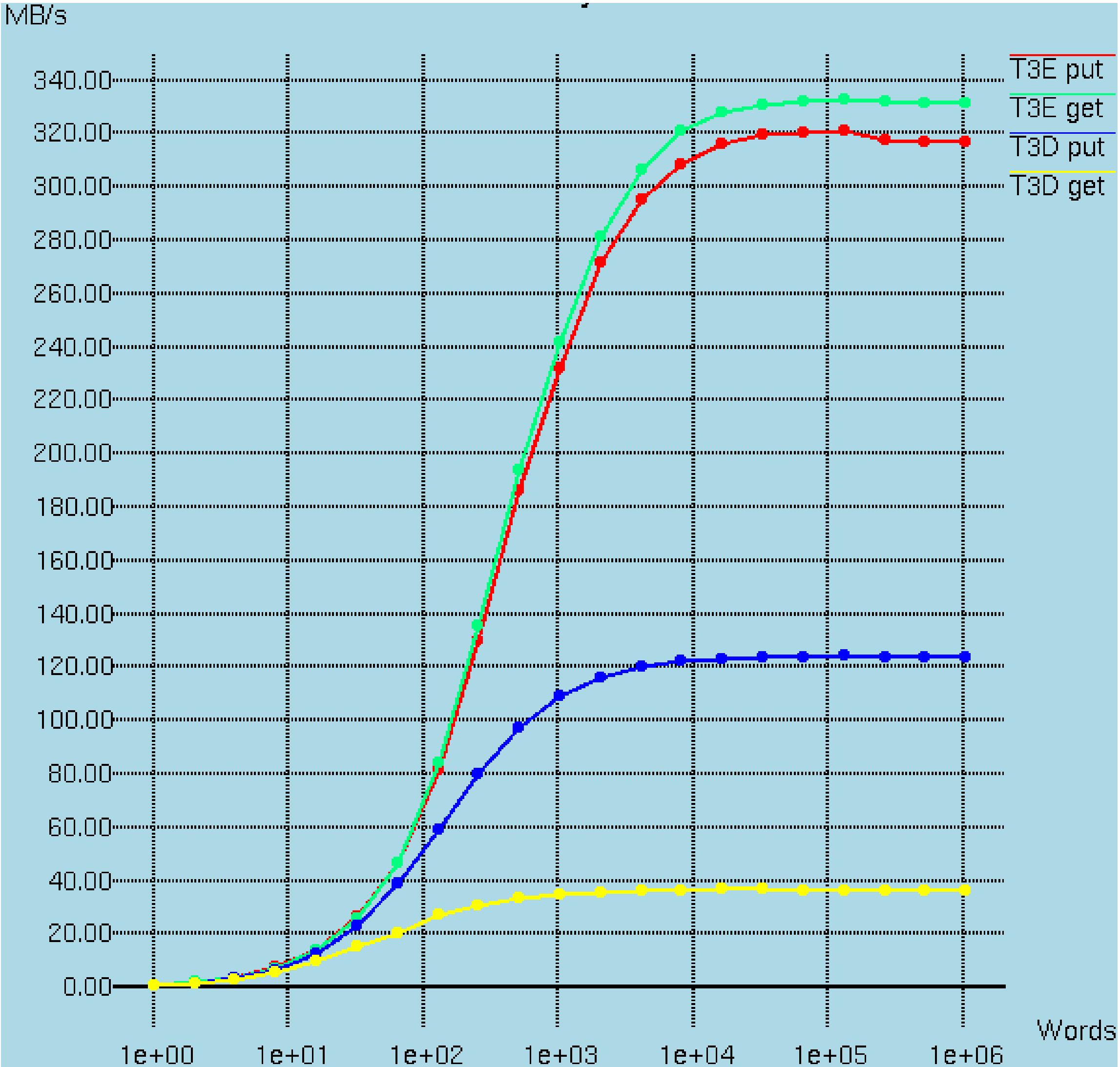
CRAY T3E - Shared Memory Broadcast

- **CRAY T3E Shared Memory Broadcast ist als Binärbaum aufgebaut**
- **Übertragungszeiten steigen mit ld n bei n PEs**

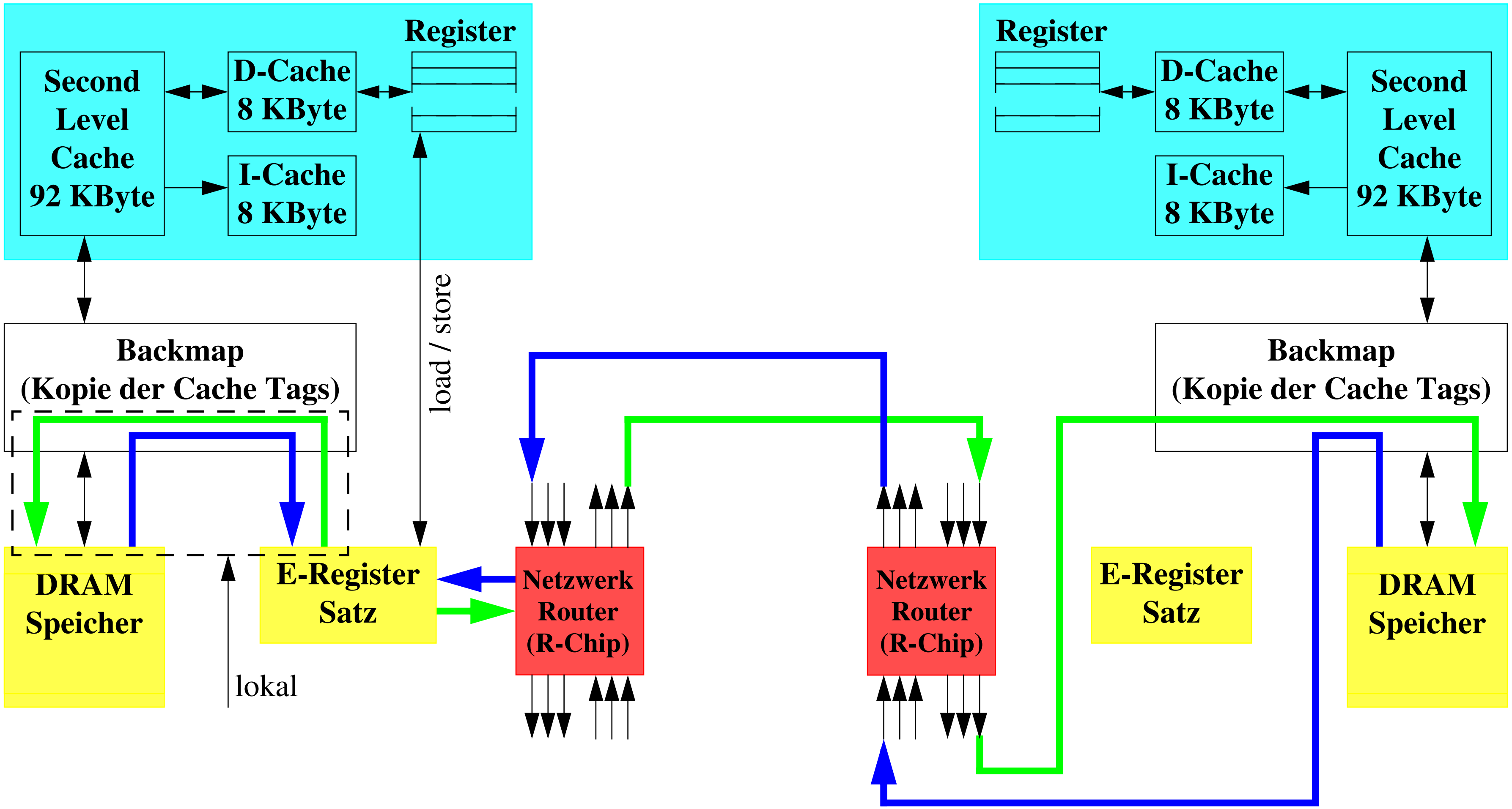


CRAY T3D / CRAY T3E - Shared Memory GET / PUT

- **CRAY T3D Shared Memory PUT deutlich schneller als GET aufgrund des Cache-Verhaltens im EV4**
- **CRAY T3E Shared Memory GET geringfügig schneller als PUT**
- **Erreichbare Bandbreiten im CRAY T3E ca. 3 mal höher gegenüber CRAY T3D**
- **Zusätzlich verfügt im CRAY T3E jeder PE über einen eigenen Netzknoten**



Adressierung des entfernten Speichers



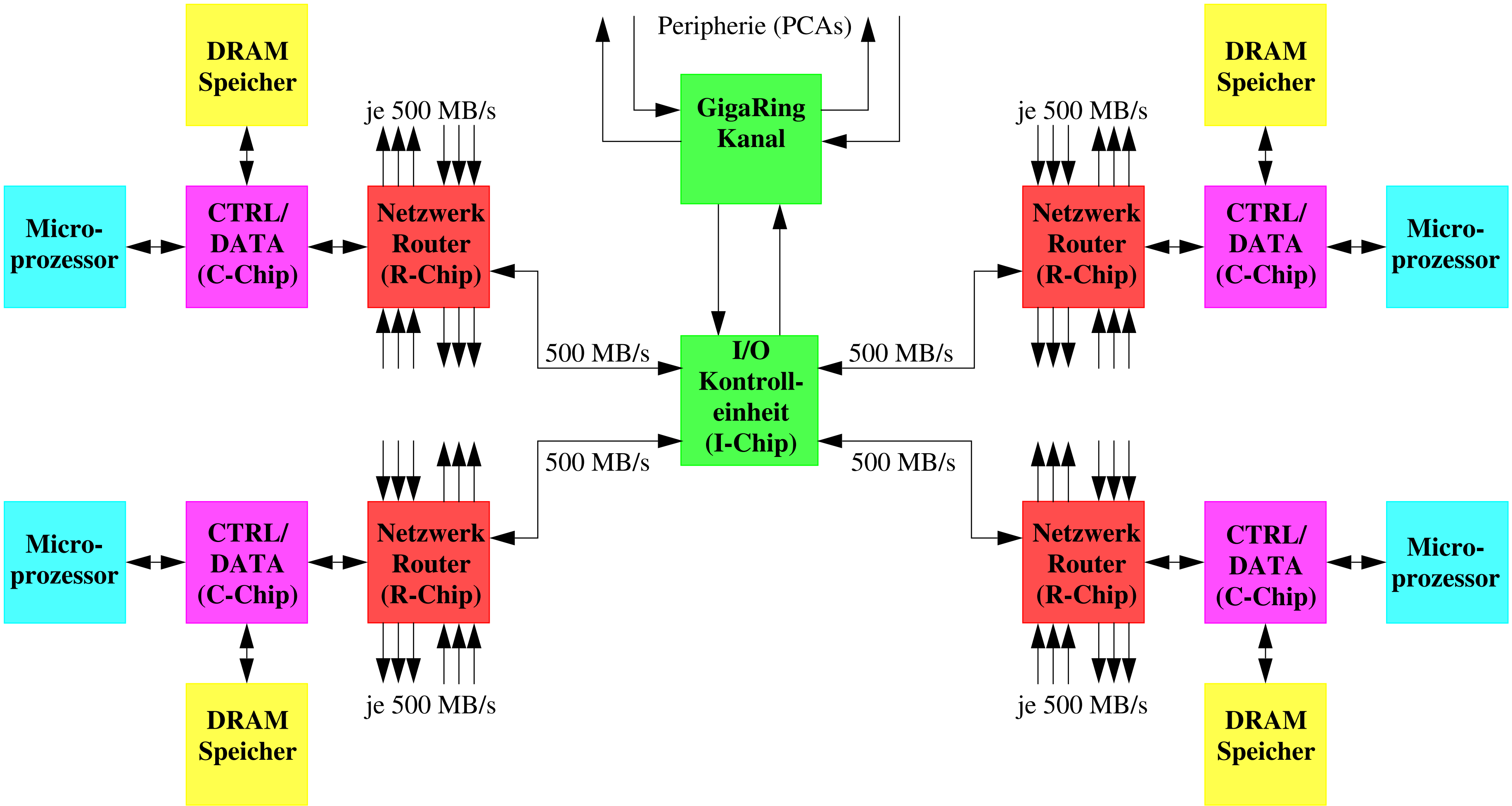
PUT-Operation transferiert vom E-Register zum DRAM Speicher

GET-Operation transferiert vom DRAM Speicher zum E-Register

CRAY T3E Message Routing

- **Das Routing der Datenpakete erfolgt durch ein Vergleich des Routing-Tags im Paket und einer Knotenadresse, die im Knoten enthalten ist**
- **Im Gegensatz zum CRAY T3D System, das mit Deltas durch fortlaufendes Dekremtieren arbeitet, werden im CRAY T3E absolute Adressen verwendet**
- **Im Knoten befindet sich eine Tabelle, die folgende Information enthält**
  - Physikalische Adresse des Zielknotens
  - Der deterministische Weg zum Zielknoten
  - Hinweis, ob adaptives Routing gestattet ist
- **Als Verweis in diese Tabelle dient die logische PE Nummer des Zielknotens**
  - Logische PE Nummer ist das, was das System als Ganzes sieht (OS)
  - Der Anwender arbeitet mit virtuellen PE Nummern

Printed Circuit Board Layout (ein I/O-Knoten je 4 PEs)



Auswirkung der Stream Buffer

- **Beispiel: mehr als 6 Ströme**

```
 do i = 1,n
  sum = sum + a(i,1)*p1 + a(i,2)*p2 + a(i,3)*p3 + a(i,4)*p4
&            + a(i,5)*p5 + a(i,6)*p6 + a(i,7)*p7 + a(i,8)*p8
 enddo
```

- **f90 -O unroll2**

Loop	n = 524288 (2^19)	CRAY T3E	
		SCACHE_D_STREAMS=1	SCACHE_D_STREAMS=0
eine Schleife	a(n,8)	6.47 MFLOPS	6.05 MFLOPS
	a(n+32,8)	65.04 MFLOPS	56.59 MFLOPS
zwei Schleifen	a(n,8)	20.78 MFLOPS	15.05 MFLOPS
	a(n+32,8)	126.75 MFLOPS	52.16 MFLOPS
zusätzlich -O split2	a(n,8)	MFLOPS	MFLOPS
	a(n+32,8)	MFLOPS	MFLOPS

Auswirkung der Stream Buffer

- **Beispiel: Horner Schema**

```
do i = 1,n
 y(i) = (((p4*x(i)+p3)*x(i)+p2)*x(i)+p1)*x(i)+p0
enddo
```

- **f90 -O unroll2**

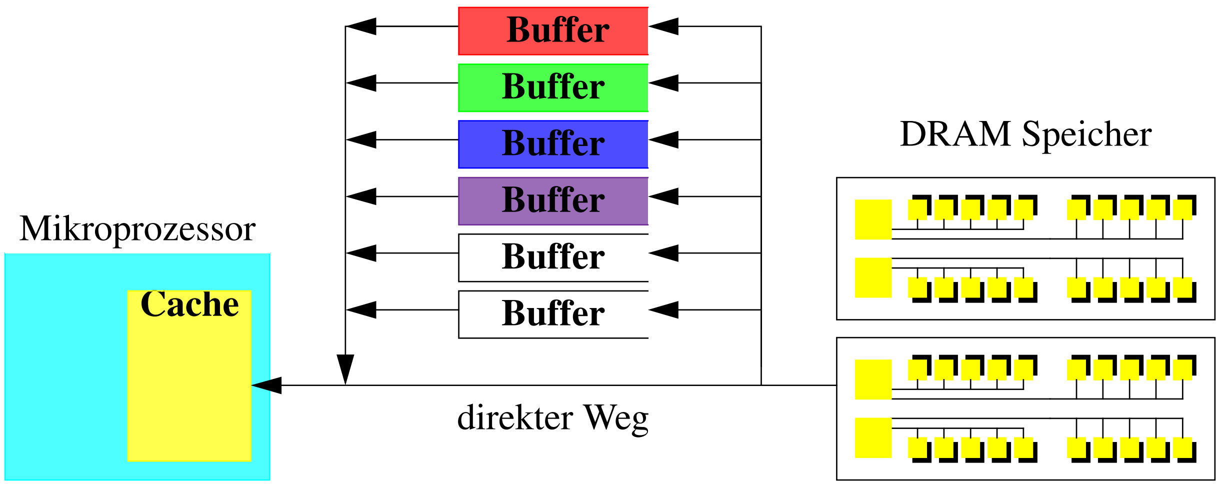
n	CRAY T3E	
	SCACHE_D_STREAMS=1	SCACHE_D_STREAMS=0
512	145.25 MFLOPS	80.00 MFLOPS
1024	186.64 MFLOPS	97.59 MFLOPS
4096	188.64 MFLOPS	85.39 MFLOPS
32768	193.48 MFLOPS	85.54 MFLOPS

Prefetch Mechanismus über Stream Buffer

- **Verkürzung der Speicherzugriffszeit durch großen externen Cache ist zu aufwendig und teuer bei massiv-parallelen Prozessorsystemen**
- **Stream Buffer erreichen, insbesondere für technisch-wissenschaftliche Anwendungen mit wesentlich geringerem Aufwand das gleiche Ziel und mehr (kein ’Herausfallen’ aus dem externen Cache)**

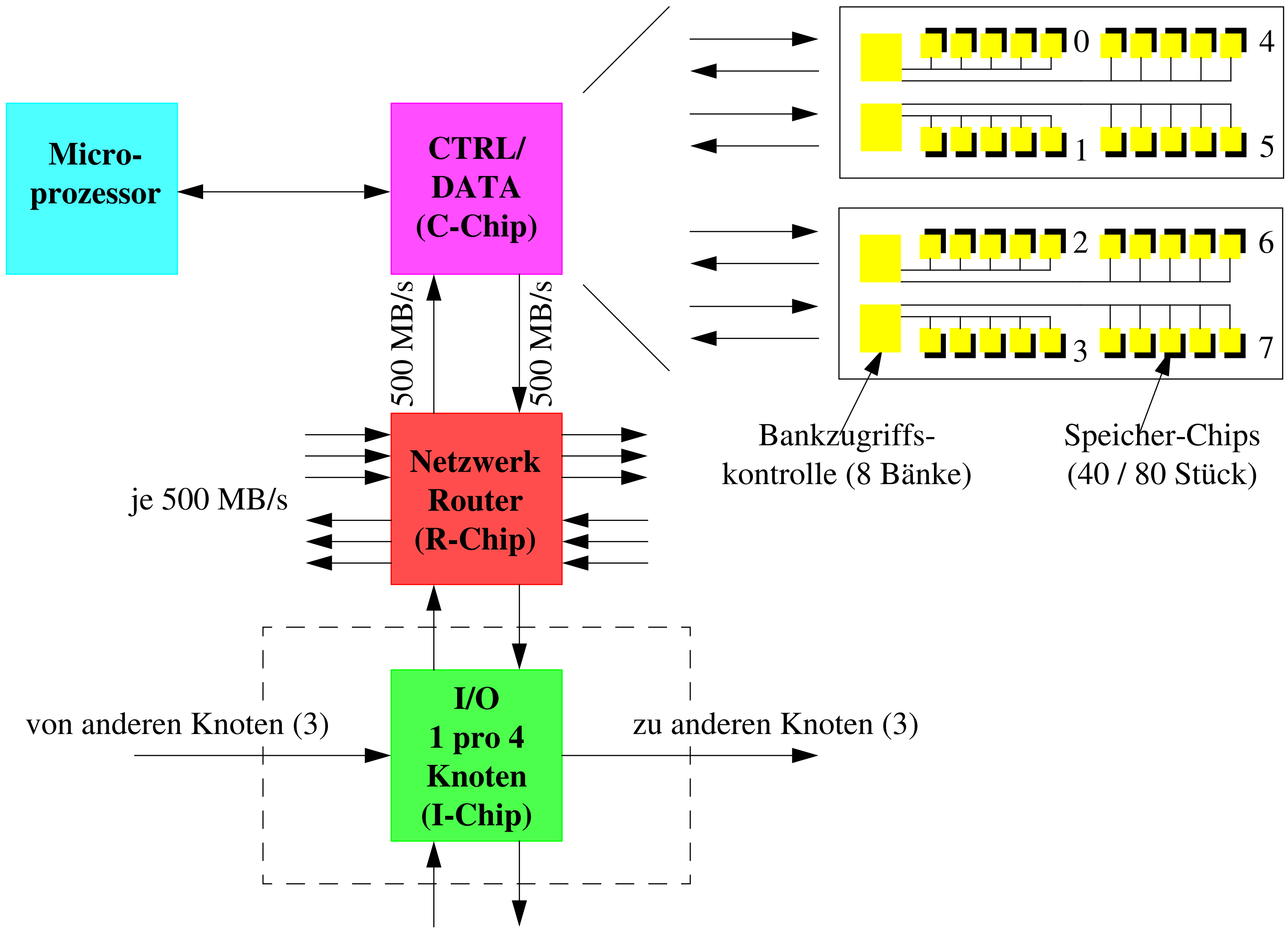
Zugriff auf aufeinanderfolgende Adressen bewirkt automatisches Vorladen (einstellbar durch Software; Shell Variable SCACHE_D_STREAMS für Daten SCACHE_I_STREAMS für Instruktionen)

6 unabhängige Streams können zeitlich verschränkt bedient werden

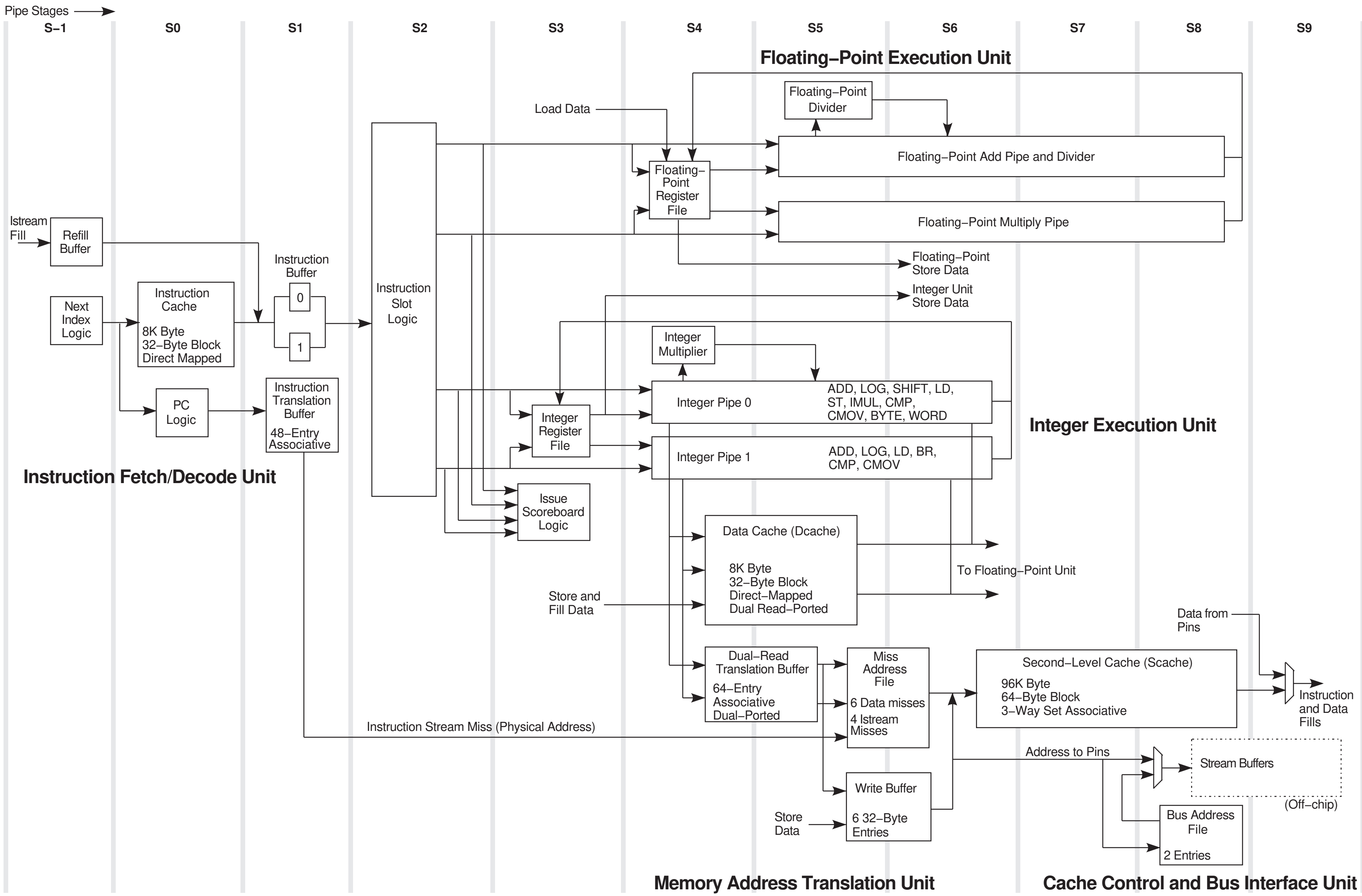



```
do i = 1,n
A(i) = B(i)+C(i)
C(i) = D(i)*E(i)
end do
```

CRAY T3E Knotenarchitektur



DEC Alpha EV5 Mikroprozessor (21164 Implementierung)



DEC Alpha EV5 Mikroprozessor (21164 Implementierung)

- **300 MHz Taktfrequenz (3.33 ns)**
- **IEEE 64 Bit Arithmetik**
- **Super-skalarer Prozessor**
  - Instruktionsrate von bis zu 4 pro Zyklus (2 Gleitkomma, 2 Integer/Logic)
  - Spitzenleistung von 1200 MIPS und 600 MFLOPS
  - Division wird auf 64 Bit durchgeführt
- **3 separate On-Chip Caches**
  - 8 KBytes Daten-Cache (D-Cache), 256 Cache-Lines zu je 4 Worten (64-Bit)
  - 8 KBytes Instruktions-Cache (I-Cache), Unterteilung wie im D-Cache
  - 96 KBytes Secondary Cache (S-Cache), unterteilt in 3 assoziative Sätze zu je 512 Cache-Lines; bedient sowohl D-Cache als auch I-Cache; Write Allocate - Write Back
- **Bandbreite des externen Datenbus beträgt ca. 1200 MB/s**

3D-Torus Topologie

- **Kurze Verbindung zwischen den Knoten**

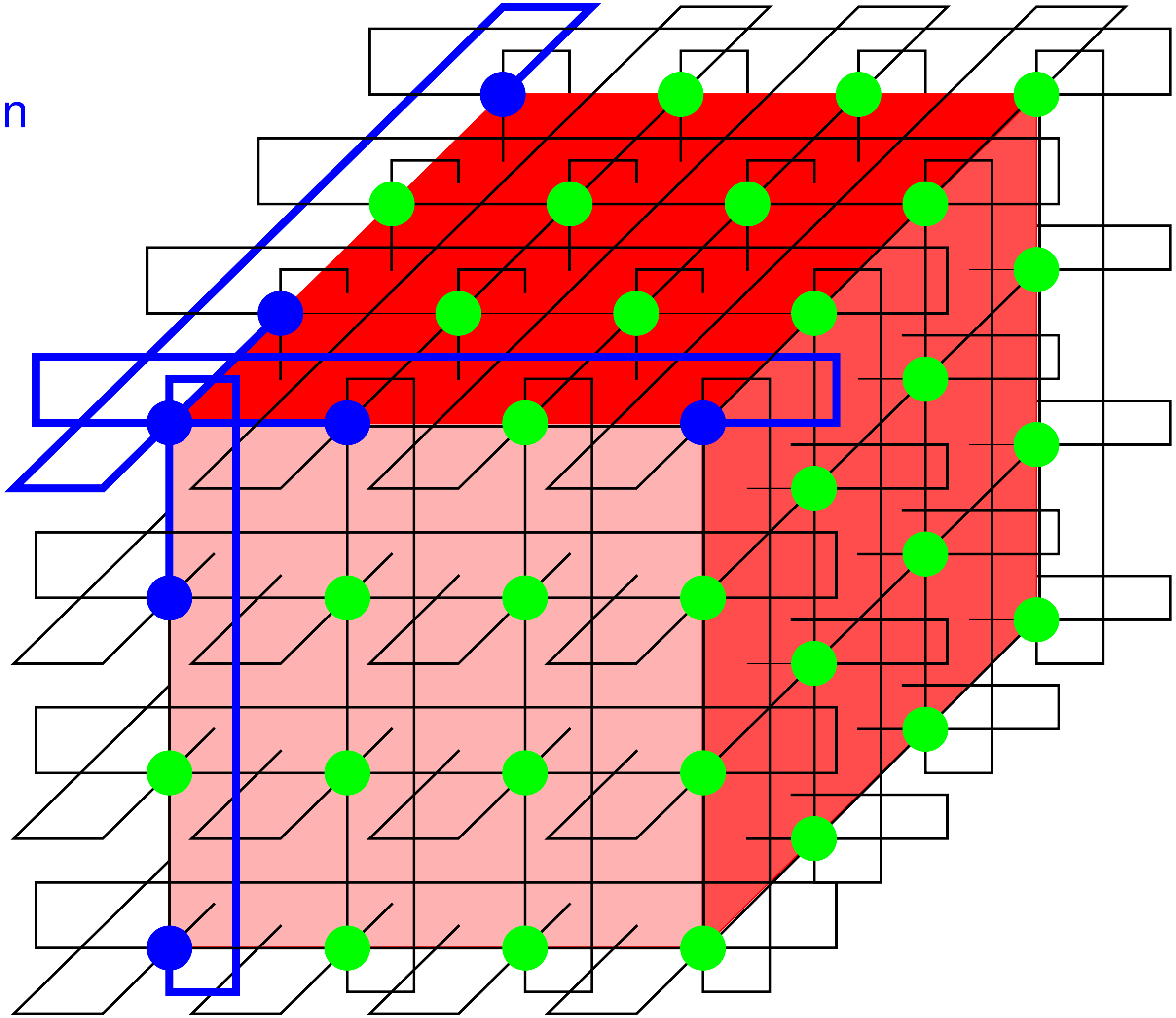
  z.B. 1024 Knoten ist ein

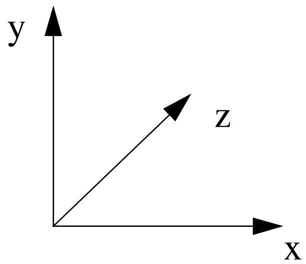
  16 x 8 x 8 Torus =>

  größte Entfernung:

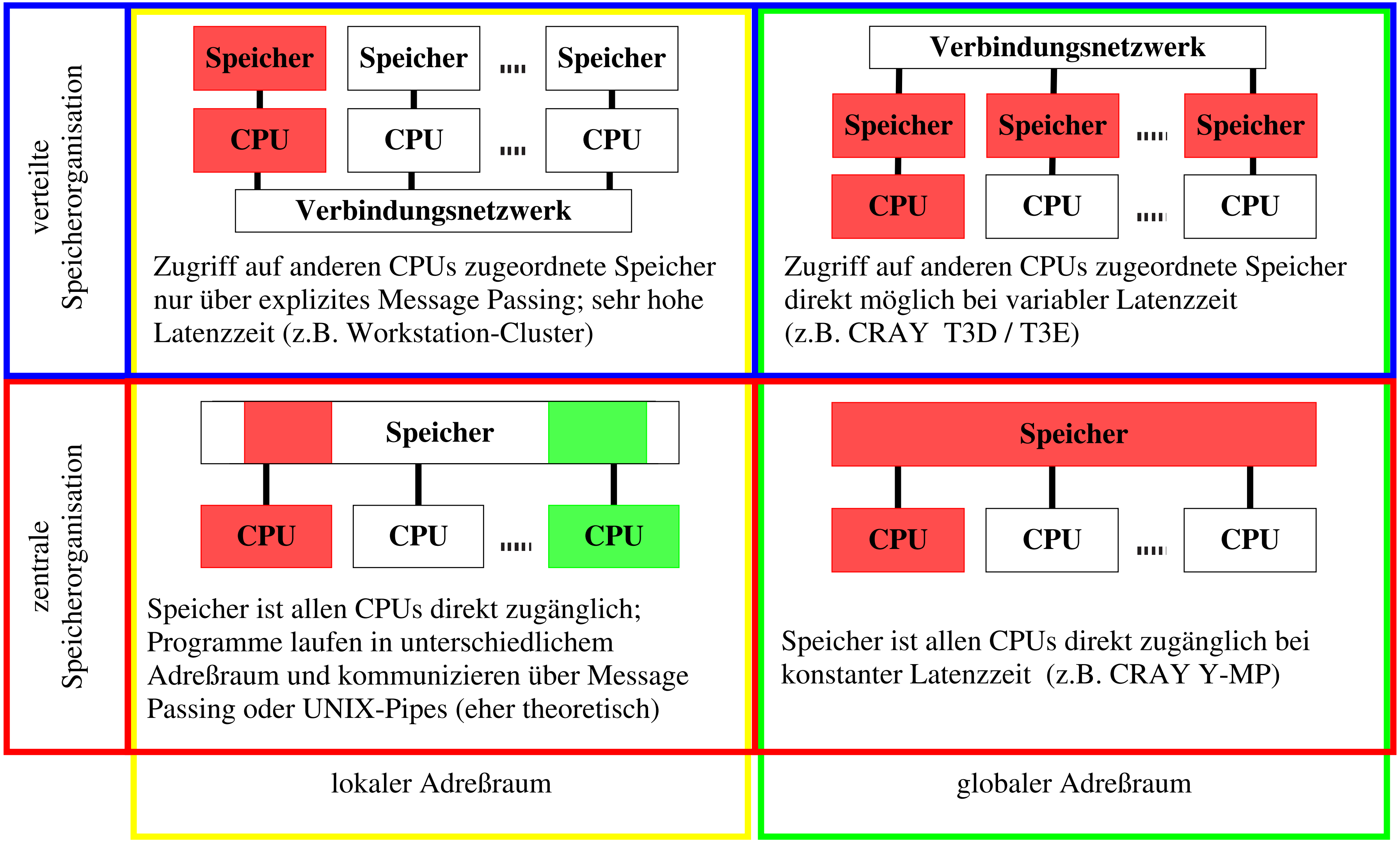
  8 + 4 + 4 = 16

- **Geringe Blockierung**
- **Hohe Bandbreite**
- **Fehlertoleranz**





Relation zwischen Speicherorganisation und Adreßraum



Weiterentwicklung des CRAY T3E aus dem CRAY T3D System

- **Ein Prozessor pro Netzwerkknoten**
- **3D Torus mit theoretischen Transferraten von ca. 500 MB/s bidirektional in jede der drei Koordinatenrichtungen ergeben eine Durchsatzleistung von 3 GB/s in einem Netzwerkknoten**
- **Speichergrößen von 128, 256 oder 512 MB pro PE**
- **Satz externer Register (E-Register) über die asynchroner Datentransfer erfolgt; ersetzt die Prefetch Queue (PFQ) und Block Transfer Engine (BLT) aus dem CRAY T3D System**
- **Cache-Kohärenz Protokoll durch Hardware**
- **Barrier/Eureka Netz ist im Kommunikationsnetzwerk integriert**
- **CRAY T3E System wird ohne Front-End betrieben (Self-Hosting); für je 16 PEs ist ein OS-Knoten vorgesehen**

Weiterentwicklung des CRAY T3E aus dem CRAY T3D System

- **CRAY T3E ist die zweite Generation des Cray Research MPP-Systems**
- **Macro-Architektur wird beibehalten**

  MIMD mit Synchronisationsmechanismen in Hardware

  Verteilter Speicher mit globalem Adreßraum ("Global Shared Memory")

  3-D Torus Verbindungsnetzwerk

  Datenverteilungs und -umverteilungsmechanismen

  Rechenknoten (Prozessor mit lokalem Speicher, Schaltelement)

  Fehlertoleranz durch redundante Knoten

  Separate I/O-Knoten
- **Skalierbar bis 1 TFLOPS Spitzenleistung**

Massiv-Paralleles Prozessorsystem

CRAY T3E