

Vergleich von XD1 (SRC) und Blue Gene (IBM)

**Seminar:
Architekturen und
Chips für
Supercomputer**

Inhalt

- Konzept und Design
 - Blue Gene
 - XD1
- Realisierung
 - Blue Gene
 - XD1
- Bewertender Vergleich
- Links und Infomaterial

Konzept und Design – Blue Gene

- Konzeptideen
- Designideen
- Folgen
- System Software
- Größe
- Anwendungsgebiete

Konzeptideen

- Gute Kosten / Leistung
 - Kombination von verschiedenen Designs (anwendungsspezifisch, universell)
- Gute Leistung / Verbrauch
- Gute Leistung / Fläche
- Allg.: low power architecture
 - Möglichst geringer Stromverbrauch

Designideen

- System on a Chip (SoC) – Design
 - Prozessor hat mittelmäßige Taktung
 - Design wird um diesen aufgebaut
- Übernahme von Ideen aus QCDSP
 - QCD on digital signal processors (10 TFLOPS)
 - Verfügte über gute Kosten/Leistung
- Verstärkung der Parallelisierbarkeit
 - Unterstützen von Anwendungen die hochgradig Parallelisierbar sind

Folgen

- Netzwerk
 - Hardwareunterstützung für manche Operationen (broadcast usw)
 - Unterstützung von sehr kleinen Paketgrößen
 - 32 bytes
- Hohe Anforderung an RAS Architektur
 - Reliability
 - Availability
 - serviceability

System Software

- Message passing
 - MPI (Message Passing Interface)
 - Programmiersprachen
 - C/C++
 - Fortran
 - Betriebssystem
 - Linux (lediglich geringfügige Kernelerweiterungen)
- Weiterhin recht günstig durch Benutzung von weit verbreiteten Standards

Größe

- Node

- Speicher: 512Mb (bis 2GB)
- Netzwerk: eigene Netzwerk Hardware

- Rack

- Maße: 0.9m*0.9m*1.9m
- Knoten: 1024 dual processor nodes
- Verbrauch: 27.5 kW

- Gesamtgröße

- $2^{16} = 65,536$ Nodes

Anwendungsgebiete

- **Spezialisierte Anwendungen**
 - Anwendungen müssen mehr Leistung durch höhere Parallelisierung erzielen
- **Energiesparend**
 - Verbraucher die hohe Leistung zu relativ wenig Verbrauch benötigen
- **Kostengünstig**
 - Durch low power und low cost Strategie relativ preiswert in Anschaffung und Gebrauch

Konzept und Design – XD1

- Konzeptideen
- Designideen
- System Software
- Größe
- Anwendungsgebiete

Konzeptideen

- Kostengünstiger Supercomputer
- Hohe Zuverlässigkeit
- Große Skalierbarkeit
- Anwendungsgebunden
 - Betriebssystem Linux
 - Weite Anwendungsbreite im Bereich OpenSource

Designideen

- Direct Connected Prozessor (DCP) Architektur
- Zuhilfenahme von FPGAs zur Steigerung der Rechenleistung
- Chip-Chip Kommunikation durch high performance interconnect

System Software

- Betriebssystem Linux
 - Direkte Unterstützung der gängigen Programmiersprachen (C/C++ usw)
 - OpenSource
 - Kostenreduktion
 - Breite Unterstützung von vielen verschiedenen HPC codes

Größe

- Chassis
 - 12 Prozessoren
 - AMD Opteron DualCore 2.2Ghz
 - 0.13m*0.58m*0.91m
- Cabinet
 - 12 Chassis
 - Performance: 1.27 TFLOPS
 - Max Memory: 1.2 TB
 - Max Speicherplatz: 18 TB
 - 26,4 kW

Anwendungsgebiete

- **Universeller Supercomputer**
 - Anwendungsorientierte Architektur (weiter gefächert als BlueGene)
 - Aber: Auf praktisch alle Anwendungen konfigurierbar
- **Preisgünstiger Supercomputer**
- **Einfachheit**
 - Wartung
 - Verwendung

Realisierung – Blue Gene

- Anwendungsorientierte Architektur
 - Anwendungen
- Prozessor und Knoten
- RAS Architektur
- High Performance Network
- Ergebnis

Anwendungsorientierte Architektur

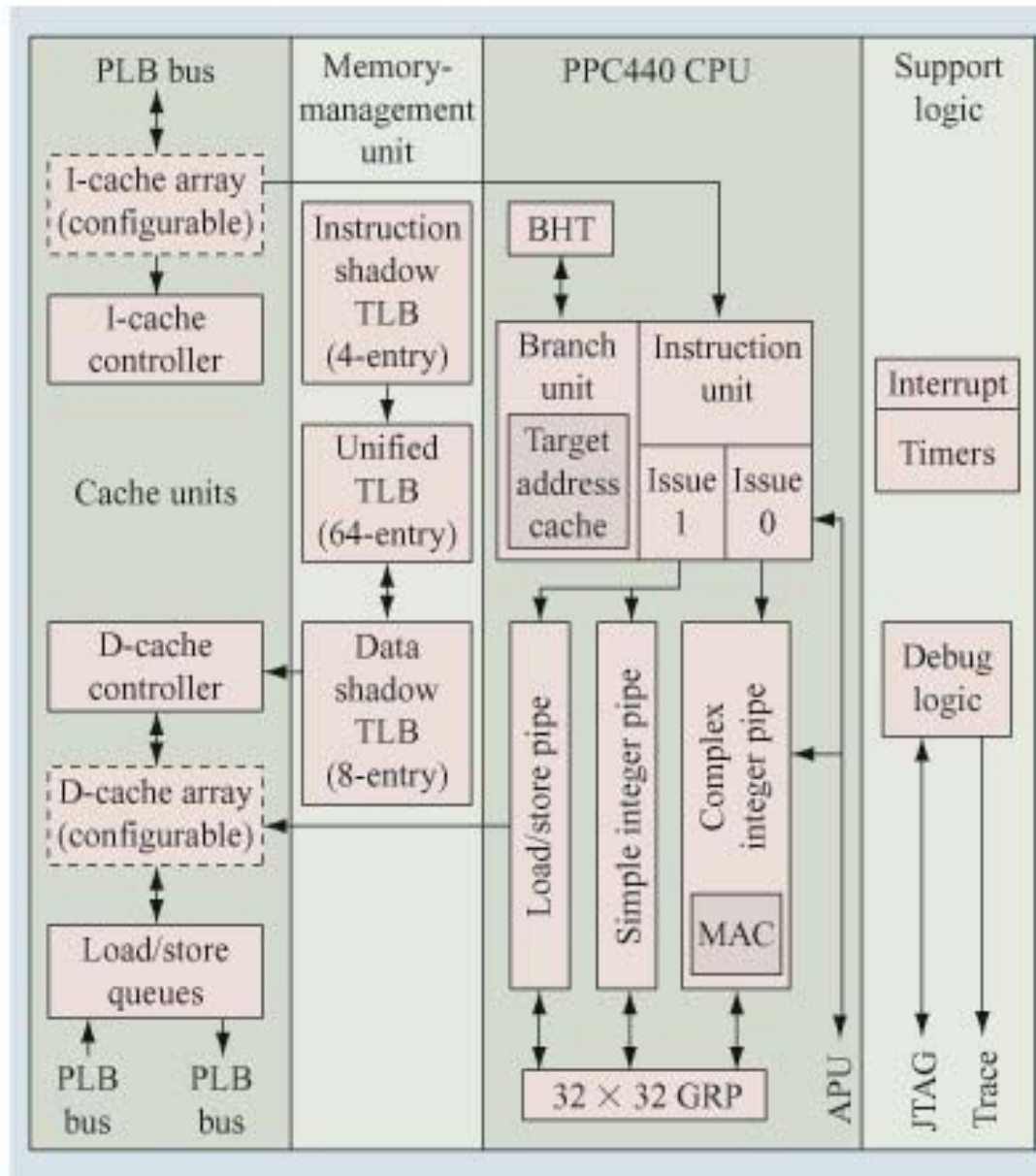
- **Vorteil:**
 - Effektive Kostendeckung für entsprechende Anwendungen
 - Hohe Leistung für diese Anwendungen
- **Nachteil:**
 - Nicht für alle Anwendungen benutzbar
 - Hohe Anforderung an Skalierbarkeit des Gesamtrechners

Anwendungen

- Simulations of physical phenomena
- Offline data analysis
- Real-time data processing
- Ersetzen von Atomtests
- Proteinfaltung
- Filmindustrie

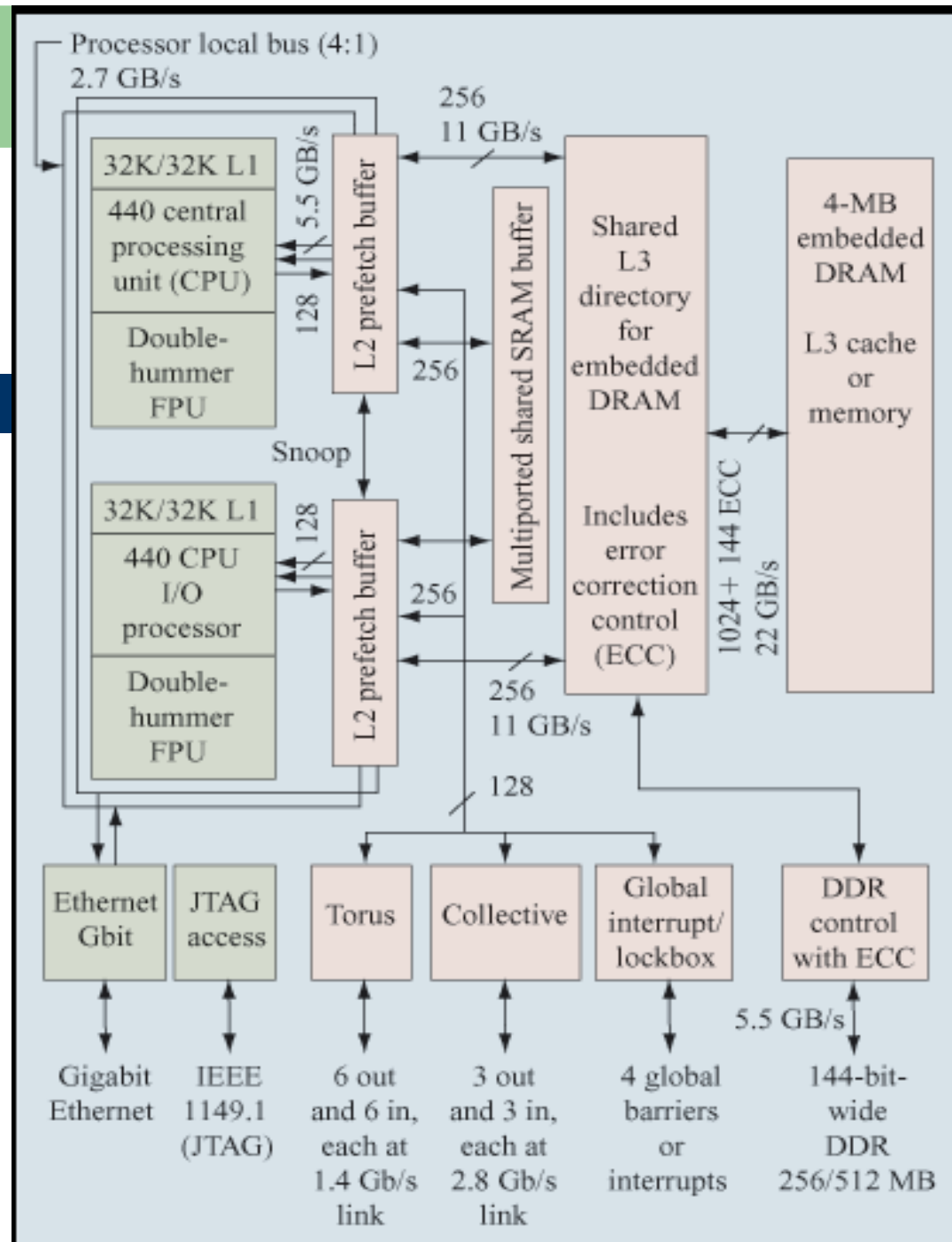
Prozessor und Knoten

- IBM PowerPC 440 (PPC440)
 - Taktrate: 700 MHz
 - Leistungsaufnahme: 1W
 - Drei unabhängige Pipelines
 - Load/Store
 - Simple Integer
 - Complex Integer
 - 32kB L1 instruction und data cache



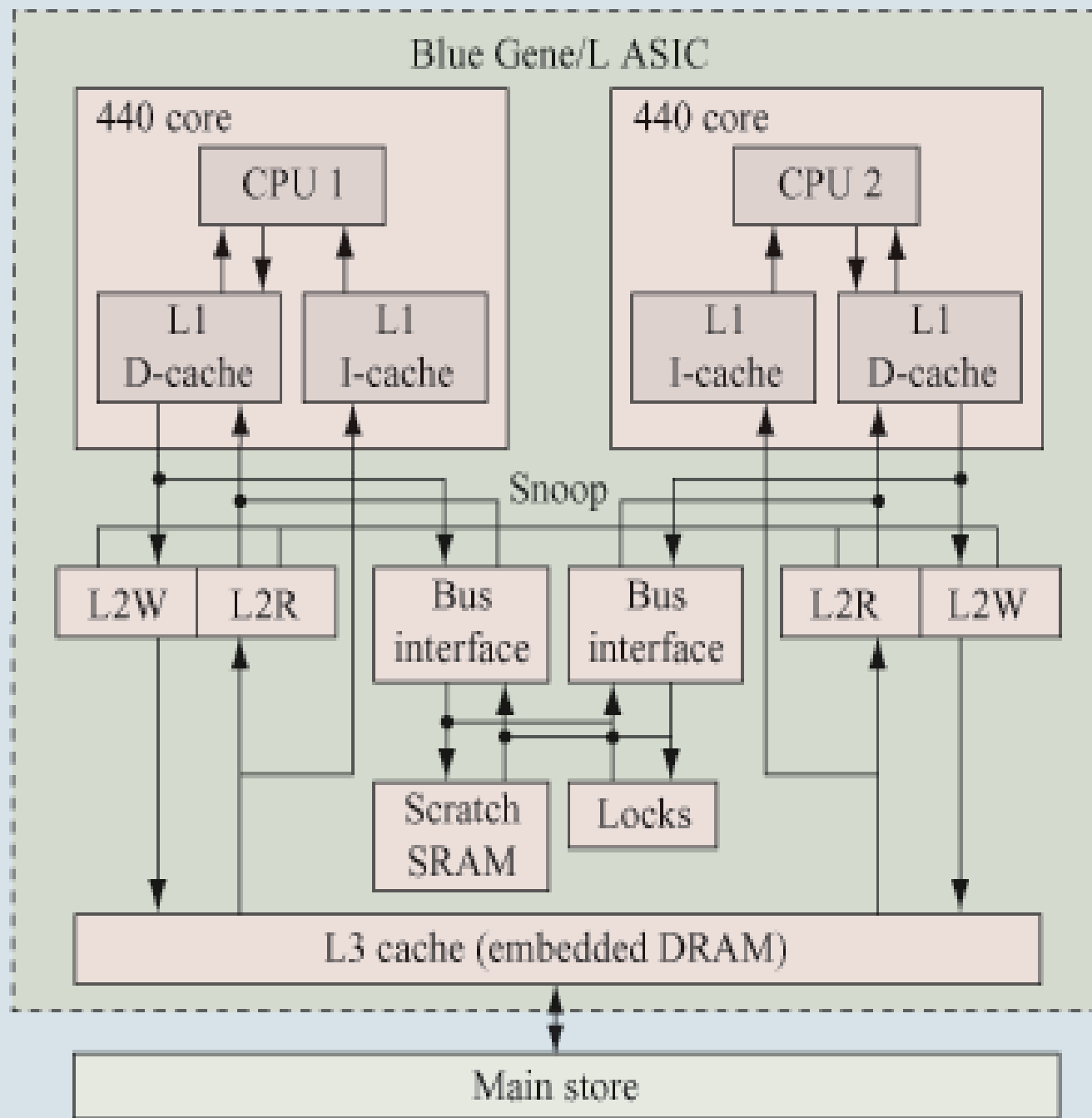
Prozessor und Knoten

- BLC ASIC
 - 2 Prozessoren
 - Erweiterte FPUs
 - Double-hammer
 - 128B Datenpfad
 - 2 FP multiply-add / cycle
 - SIMD Erweiterungen



Prozessor und Knoten

- Speicher
 - On-chip L1 cache
 - Kleiner L2 cache
 - Großer L3 cache
 - high bandwidth
/ low latency



RAS Architektur

- System level
 - Wenige verschiedene Komponenten
 - Weniger Fehlerquellen
 - Redundanz
 - Stromversorgung
 - Lüfter
 - Einfaches Clock Netzwerk
 - Zusätzliche Abschirmung der Kabel zwischen den einzelnen Racks

RAS Architektur

- BLC ASIC
 - ECC (error checking and correction) Protokoll
 - Mehrere Fehler-Detektoren im Netzwerk
 - Paketneusendung durch Hardware
 - Self-checking Algorithmen
 - Aufspüren von „illegalen“ Zuständen
 - Fehler werden gespeichert und protokolliert
 - Dient zur Fehlervorhersage und damit Fehlervermeidung

RAS Architektur

- Isolation
 - CRC Detektoren an jedem Ein-/Ausgang der Knoten
 - Vermeidet Fehlerweiterleitung durch defekte Knoten
 - Checksummmen werden mit jedem Paket gesendet
 - Auffinden fehlerhafter Knoten
 - Halt System
 - System kann komplett gestoppt werden

High Performance Netzwerk

- Netzwerk geteilt in 5 einzelne Netzwerke
 - Gigabit Ethernet
 - File system access
 - Fast Ethernet (100MBit) and JTAG
 - Debugging, diagnostics, initialization
 - Drei high-bandwidth, low-latency Netzwerke
 - Torus Netzwerk
 - Collective Netzwerk
 - Barrier Netzwerk

High Performance Netzwerk

- Torus Network
 - Drei-dimensionale Vernetzung der einzelnen Knoten
 - 6-fache Anschlussmöglichkeit
 - Nearest neighbor network
 - Cut-through traffic
 - Teilen des Netzwerkes für weiter entfernte Knoten
 - 1.4Gb/s nearest neighbor links (beide Richtungen)
 - Bandbreite 2.1GB/s bei 100ns Latenz
 - Maximale Entfernung 64 Hops → max Latenz 6.4 μ s

High Performance Netzwerk

- Collective Network
 - Bandweite 4bits / processor cycle (2.8Gb/s)
 - Unterstützt durch zusätzliche Hardware Integer Operationen
 - Min, max, sum
 - Bitweises OR, AND, XOR
 - Ermöglicht broadcast zusätzlich zu Datenverkehr auf dem Torus Netzwerk
 - Unterstützt virtuelle Unternezwerke
 - Benutzer Partitionen

High Performance Netzwerk

- Barrier Network
 - Bildet „globales“ OR
 - AND durch invertierte Logic
 - OR gilt als Interrupt
 - System kann angehalten werden in $\sim 1.5\mu\text{s}$
 - AND gilt als Barriere
 - Unterstützt Partitionen
 - Dient vor allem zum Zählen und Sortieren von Knoten

High Performance Netzwerk

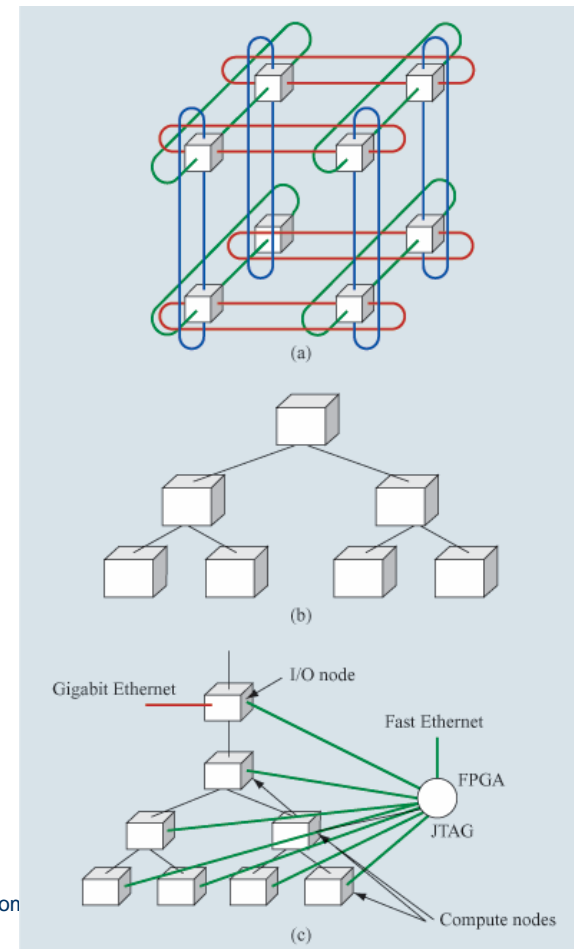
- Control System Network
 - Dient zur Initialisierung und Wartung von Endpunkten (Thermometer, Lüfter, Leuchtdioden, Netzteile usw.)
 - Wird von einem sog. „service node“ gesteuert
 - FPGA
 - Umwandlung der Ethernet Pakete in JTAG zum Debuggen und laden von Programmen
 - Zugriff auf jedes einzelne Register in jedem Knoten (IBM RiscWatch)

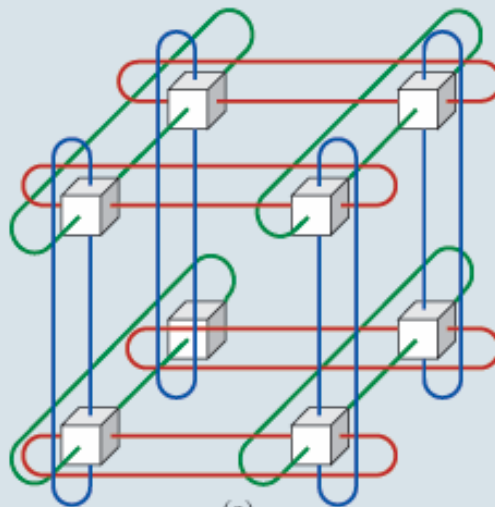
High Performance Netzwerk

- Gigabit Ethernet Network
 - Verwaltet und steuert I/O Datenzugriffe
 - Anzahl der I/O Knoten konfigurierbar
 - Max IO-to compute node Rate 1:8
 - Bei vollständigem Blue Gene:
 - Rate 1:64 bei 64 Racks
 - 1024 IO-Nodes
 - Datenrate von mehr als einem Terabit / sec

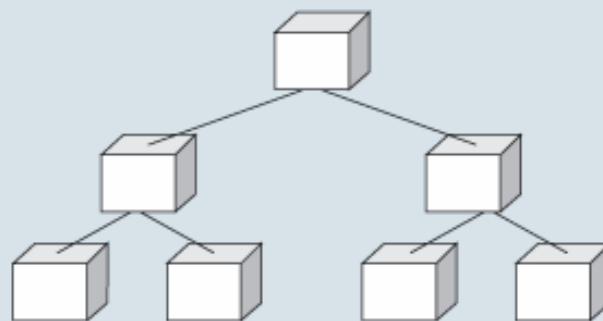
High Performance Netzwerk

- (a) 2 x 2 x 2 torus network
- (b) einfaches collective network
- (c) control system network und Gigabit Ethernet Anbindung

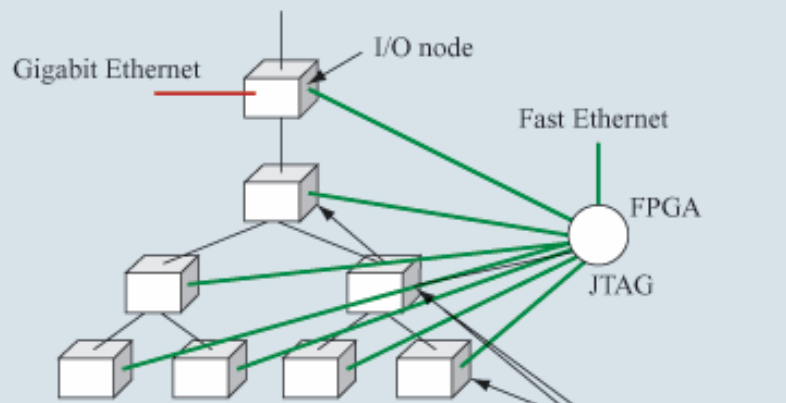




(a)



(b)



(c)

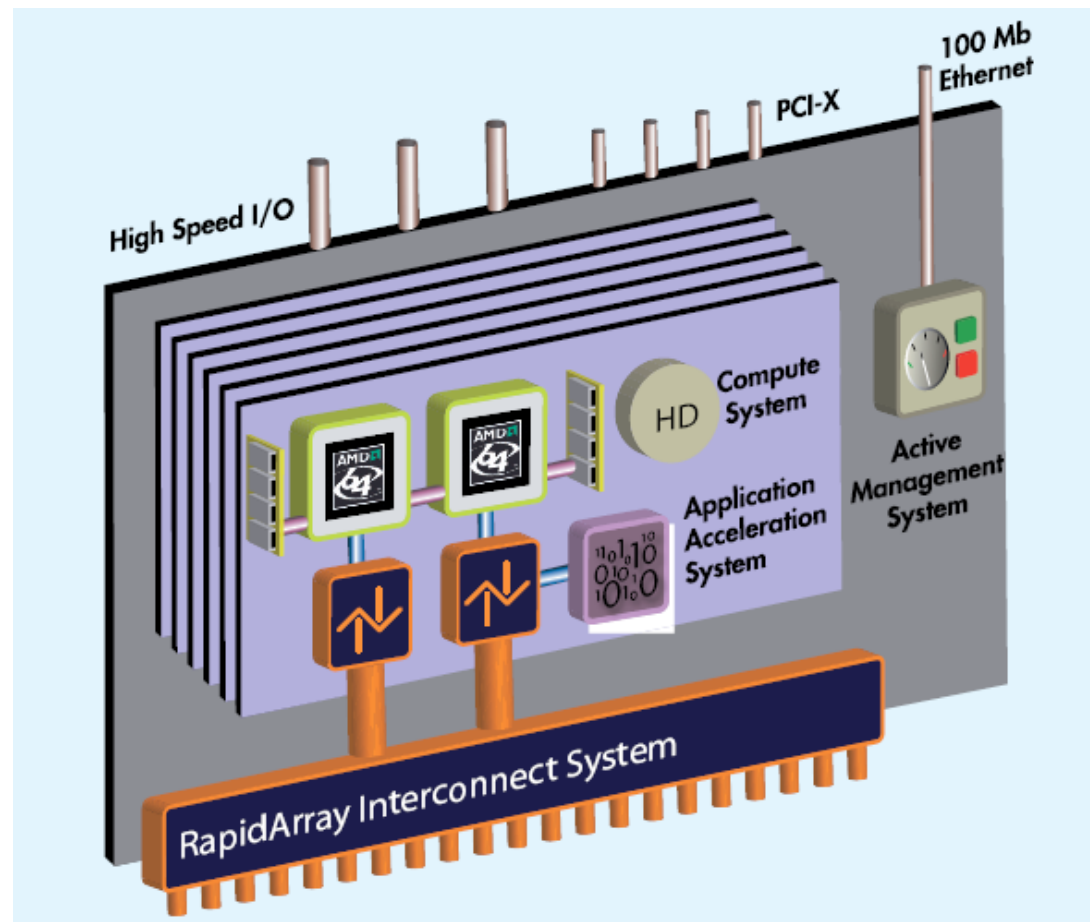
Ergebnis



Realisierung – XD1

- Prozessor
- RapidArray™ Interconnect
- Active Management
- Application Acceleration
- Ergebnis

Schematische Darstellung



Prozessor

- AMD Opteron DualCore
 - Taktrate: 2.2GHz
 - Bitrate: 64bit
 - Cache pro Kern
 - 64kB L1 Cache (instruction und data)
 - 1MB L2 Cache

Compute Environment

- High Performance Linux System
 - HPC enhanced linux kernel 2.6.5
- Dateisysteme
 - NFS v2/3
 - ReiserFS
 - Lustre Global Parallel File System
- Parallel Processing
 - MPI 1.2
 - Sockets Direct Protokoll (SDP) für TCP/IP Beschleunigung

RapidArray™ Interconnect

- Direkte Verbindung der einzelnen Prozessoren
 - High-speed – low-latency
- RapidArray Communication Processors
 - Übernehmen die Kommunikation zwischen den einzelnen Opteron Prozessoren
 - Ermöglichen beschleunigte Kommunikation
 - Verbessern die Rechenleistung der einzelnen Opterons

RapidArray™ Interconnect

- RapidArray Embedded Switching Fabric
 - 2GB/s links zwischen einzelnen Knoten und Chassis
- RapidArray Communications Libraries
 - Umgehen den Linux Kernel
 - Verbessern die Kommunikation durch eigenen Scheduler
- Ersetzt ein Ethernet zwischen den Knoten
 - Umgeht somit TCP/IP overheads

Active Management

- Management Prozessor in jedem Chassis
 - Graphische Oberfläche
 - Überwachung und Wartung
 - Single System Command Control
 - Bearbeitung von Partitionen nicht von Einzelkomponenten
 - Self-Healing
 - Fehlererkennung und Vorhersage
 - Automatische Fehlerbehebung

Active Management



Application Acceleration

- Realisiert als FPGA
 - Xilinx Virtex-4 FPGA
 - Programmierbar
 - Beschleunigung von Schlüssel-Algorithmen
 - Fungieren als Co-Prozessor
 - Berechnungen können vom Opteron Prozessor ausgelagert werden
 - Unabhängige Berechnung
 - Standard Software
 - Einfache Programmierung

Application Acceleration

- 16MB QDR Ram
- 3.2 GB/s interconnect
- Direkte Manipulation durch den Prozessor
 - Spezielle Bibliotheken
 - On-the-fly Programmierung
- Paralleles Co-Prozessor Netzwerk
 - Unabhängige Kommunikation
 - Bis zu 6 FPGAs pro Chassis

Ergebnis



Bewertender Vergleich

- Leistung
- Kosten
- Vorteile/Nachteile
 - BlueGene
 - XD1

Leistung

- Rechenleistung
 - BlueGene
 - Bis ~300 TFLOPS
 - XD1
 - Bis ~1.27 TFLOPS pro Cabinet
- Energieverbrauch
 - BlueGene
 - 211 MFLOPS / Watt
 - XD1
 - 48 MFLOPS / Watt

Kosten

- BlueGene
 - Bis ~100,000,000 Dollar
- XD1
 - 100,000 – 2,000,000 Dollar

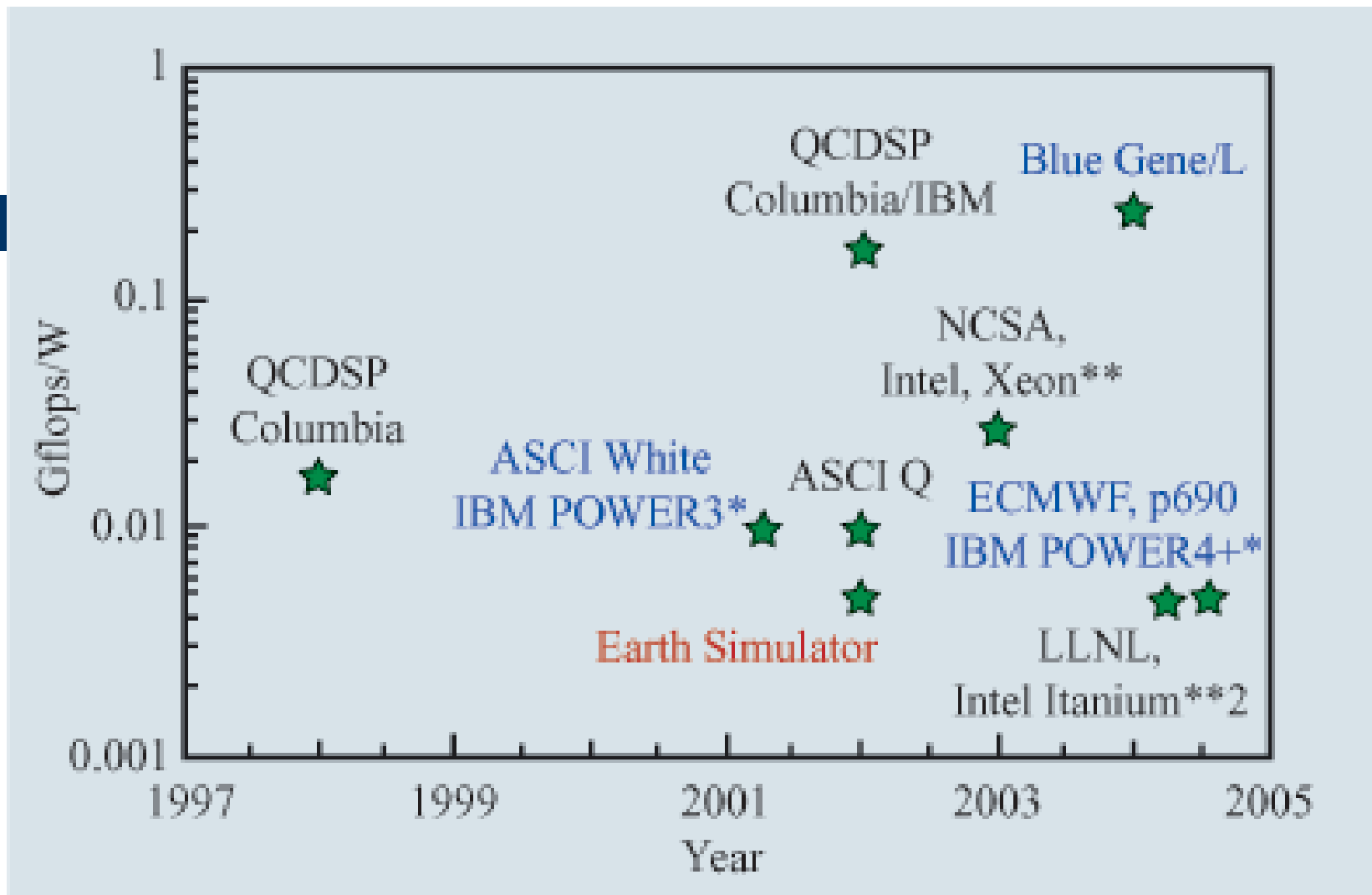
BlueGene

- Vorteile:

- Sehr große Rechenleistung
- Rechenleistung extrem skalierbar
- Guter Stromverbrauch
- Spezialisiert

- Nachteile:

- Spezialisiert
- Günstig geplant aber dennoch sehr teuer
- Gute aber komplizierte Wartung



XD1

- Vorteile:
 - Gute Rechenleistung für wenig Geld
 - Kompaktes Design
 - Skalierbar
 - An nahezu jede Anwendung anzupassen
 - Benutzerfreundlich
- Nachteile:
 - Leistung/Watt nicht optimal
 - Gesamtleistung

Links und Infomaterial

- <http://www.research.ibm.com/journal/rd/492/gara.html>
- http://www.cray.com/downloads/Cray_XD1_Datasheet.pdf
- http://www.xilinx.com/products/silicon_solutions/fpgas/virtex/virtex4/index.htm
- <http://www.itec.uni-karlsruhe.de/capp/teaching/pap/ws06/pap06-01.pdf>
- <http://www.tecchannel.de/server/hardware/402401/>