Systemsoftware II

19. Trends

19.1

Größer, Schneller und Teurer

109.951.162.776 Bytes in 1057 Sekunden sortiert

http://www.austin.ibm.com/

Ordnung auch in der größten Unordnung

Die Aufgabe

- 10 Milliarden Datensätze
- Datensatz
  - 10 Byte Schlüssel (zufällig verteilt)
  - 90 Byte Zusatzdaten (Erhöhen E/A-Anforderung)
- Gesamtgröße: 1 Terabyte (1613 randvolle CD-ROM)
- Eingabe und Ergebnis auf Externspeicher
- Zeit inkl. Programmstart etc.

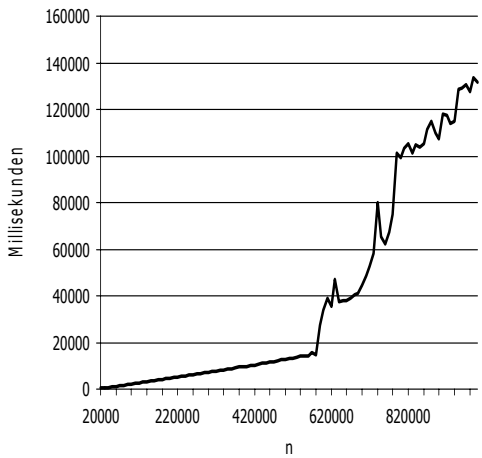
19.4

Na und?

- Enorme Speichermenge
  - 1 Tbyte = 69 Platten zu 15 Gbyte
  - 17.000 DM bei 260 DM pro 15 G-Platte (November 1999)
- Viele Platten !!!
- Bandbreite zur Platte kritisch
  - Einmal 1 Tbyte lesen: 29 Stunden und 7 Minuten
    - Annahme: 1 Platte, 10 Mbyte/s Durchsatz
  - 1 Tbyte in 17 Minuten lesen: 102 Platten parallel
    - Bandbreitenbegrenzung Platte-Rechner (Bus)
- Viele Platten parallel an vielen Rechnern !!!

19.5

Sequentielles Sortieren 1



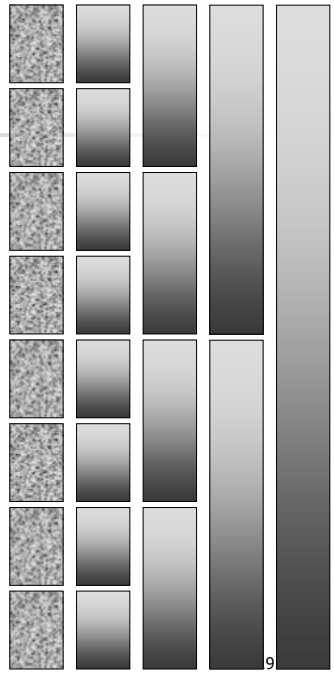
- Daten im Speicher
  - Zugriffszeit Nanosekunden
- Beispiel: Quicksort
- Messung
  - Pentium II, 450 MHz
  - Genügend Hauptspeicher
    - 83 Jahre (linear) (1 Prozessor)
    - 15 Tage (2000 Prozessoren)
  - Komplexität:

$$O(n \cdot \log n)$$

19.6

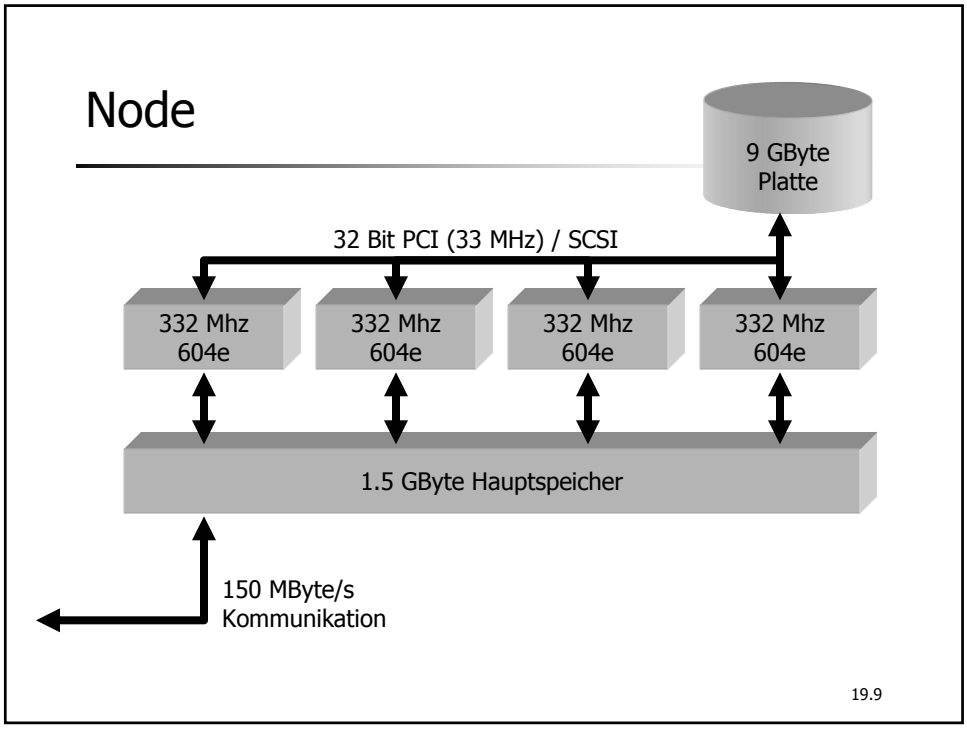
Sequentielles Sortieren 2

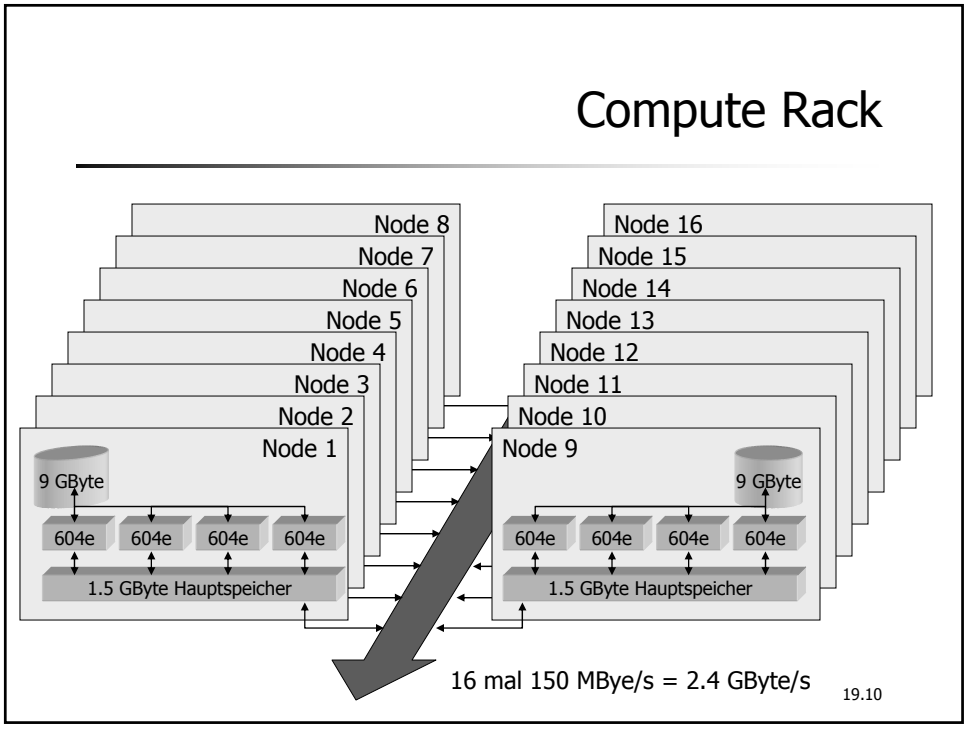
- Hauptspeicher zu klein
  - Daten befinden sich auf Platten
  - Zugriffszeit Millisekunden
  - Daten möglichst selten bewegen
- Beispiel: Mergesort
- Parallelitätsgrad sinkt gegen Ende

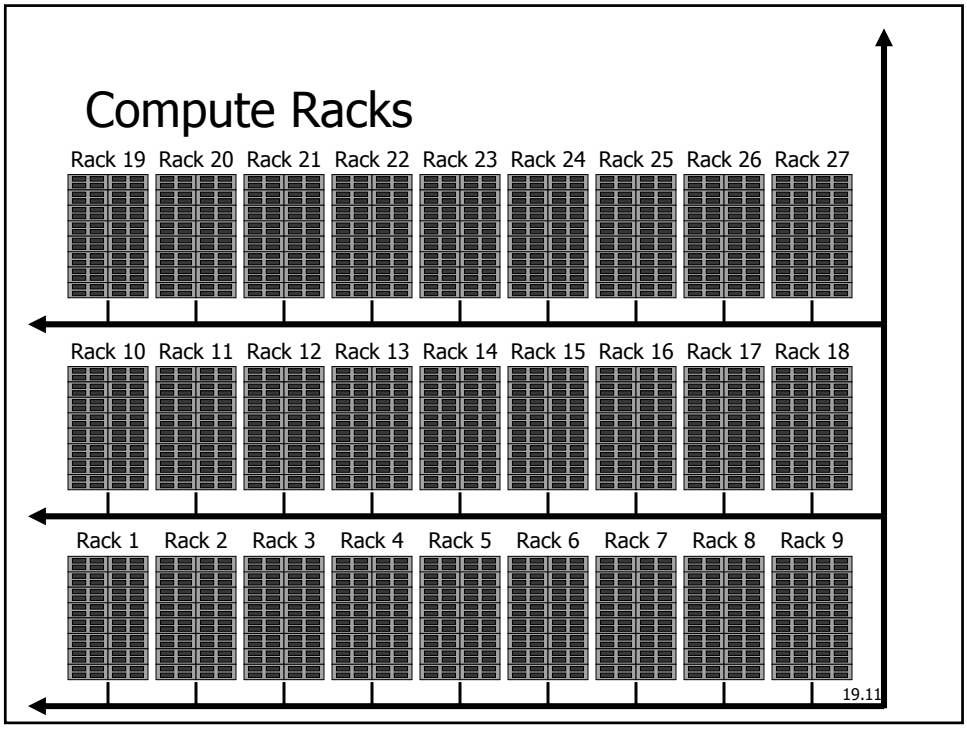


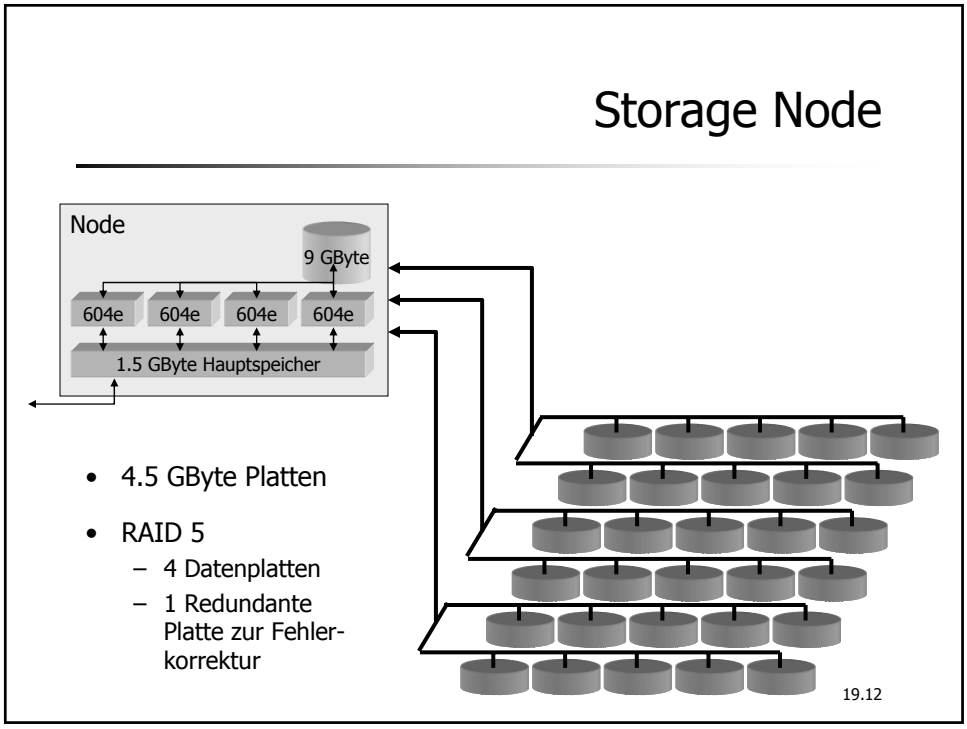
9

Der Weltrekordrechner

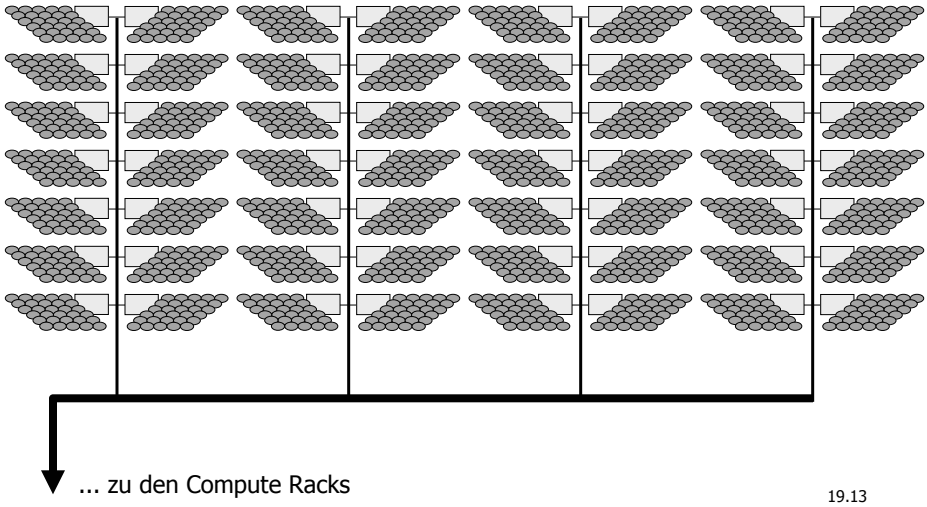

Node
9 GByte Platte
32 Bit PCI (33 MHz) / SCSI
332 Mhz 604e
332 Mhz 604e
332 Mhz 604e
332 Mhz 604e
1.5 GByte Hauptspeicher
150 MByte/s Kommunikation
19.9


Compute Rack
Node 8
Node 7
Node 6
Node 5
Node 4
Node 3
Node 2
Node 1
Node 16
Node 15
Node 14
Node 13
Node 12
Node 11
Node 10
Node 9
9 GByte
9 GByte
604e
604e
604e
604e
604e
604e
604e
604e
1.5 GByte Hauptspeicher
1.5 GByte Hauptspeicher
16 mal 150 MBye/s = 2.4 GByte/s
19.10


Compute Racks
Rack 19 Rack 20 Rack 21 Rack 22 Rack 23 Rack 24 Rack 25 Rack 26 Rack 27
Rack 10 Rack 11 Rack 12 Rack 13 Rack 14 Rack 15 Rack 16 Rack 17 Rack 18
Rack 1 Rack 2 Rack 3 Rack 4 Rack 5 Rack 6 Rack 7 Rack 8 Rack 9
19.11


Storage Node
Node
9 GByte
604e 604e 604e 604e
1.5 GByte Hauptspeicher
• 4.5 GByte Platten
• RAID 5
– 4 Datenplatten
– 1 Redundante Platte zur Fehler-korrektur
19.12

56 Storage Nodes



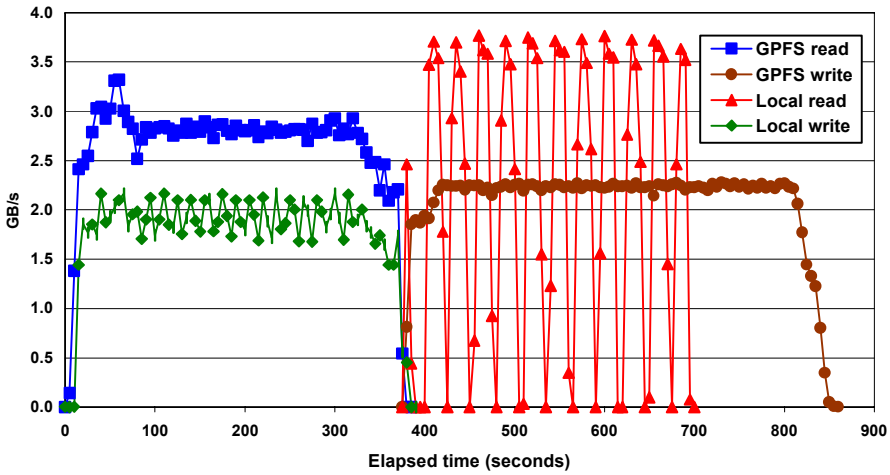
19.13

In Zahlen

- Prozessoren
  - 27 mal 16 mal 4 = 1728 Arbeitsprozessoren
  - 56 mal 4 = 224 E/A-Prozessoren
  - 1952 Prozessoren insgesamt
- Hauptspeicher
  - (27 mal 16 + 56) mal 1.5 GByte = 732 GByte
  - 1.7 Millionen DM bei 300 DM pro 128 MByte
- Lokale Platten
  - (27 mal 16 + 56) mal 9 GByte = 4.2 TByte
  - 122000 DM bei 250 DM pro 9 GByte Platte
- RAID-System
  - 56 mal 6 mal 5 Platten = 1680 Platten
  - 4.5 GByte pro Platte = 7.38 TByte

19.14

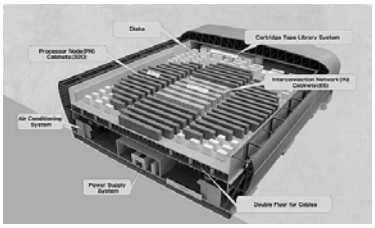
Ergebnis bzgl. E/A-Leistung



19.15

... noch größer

- Earth Simulator, Japan
- Kenngrößen
  - 5,120 (640 8-way nodes) 500 MHz NEC CPUs
  - 8 GFLOPS per CPU (41 TFLOPS total)
  - 2 GB (4 512 MB FPLRAM modules) per CPU (10 TB total)
  - shared memory inside the node
  - 640 × 640 crossbar switch between the nodes
  - 16 GB/s inter-node bandwidth
  - 20 kVA power consumption per node

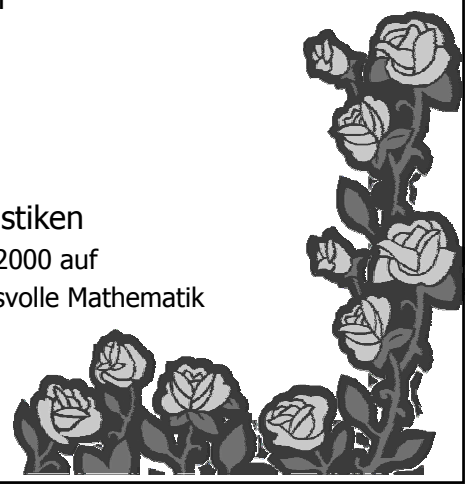


19.2

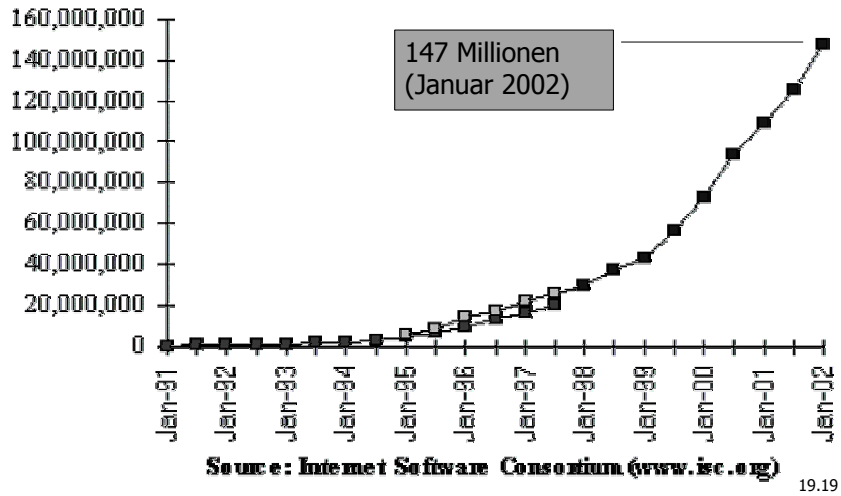
The Grid

Es wächst und wächst …

- Internet wird kontinuierlich
  - Größer
  - Schneller
  - Dichter
  - Vielschichtiger
- Überraschend wenige Statistiken
  - Viele WWW-Sites hören ca. 2000 auf
  - Gründe: zu groß?, anspruchsvolle Mathematik



Größe des Internets



19.19

Das Internet von Gestern

- Dominanz Client/Server
  - Vergleichsweise wenige zentrale Server
  - Ausgeprägte Asymmetrie
    - Client: Dump Client, Network Computer
    - Server: Dicke Sites (Cluster)
  - Hohe Last im „Inneren" des Internets
- Zentralistisches Weltbild
  - Wesentliche Managementstrukturen aus den 70er Jahren
  - Überschaubare, verwaltbare Rechneranzahl

19.20

Das Internet von Heute (1)

- Verbesserte Kommunikationsqualität
  - Bessere Netztechnologien
  - Höhere Übertragungsraten
  - Bessere Latenz (Kurze Ping-Zeiten)
- Vision „Single System Image" anwendbar auf räumlich immer größere Rechnernetze
  - Hinreichend gute Zeitsynchronisation (≤ 1 ms)

19.21

Das Internet von Heute (2)

- Wirklich JEDER (☹) ist mittlerweile im Internet
  - Flate Rate, Always On - Jeder Rechner ist Server
  - Asymmetrische Up- und Downlinks
  - Dynamische IP-Adressen
- Trend vom Dump Client zum Peer
  - Wissenschaftliche Herausforderung: Skalierbare, symmetrische und selbstorganisierende Lösungen
- Demokratisierungsbewegung (68er des Internets)
  - Nähe zur politischen „Open Source"-Bewegung
  - Peer-to-Peer Bewegung
  - Schutz, Sicherheit, Privatsphäre, Anonymität

19.22

Wozu Grids?

- Distributed Supercomputing
  - Proteinfaltung, Simulationen in Physik, Chemie, etc.
  - Verteilte interaktive Simulationen (DIS, Spiele ☺)
- High Throughput
  - Scheduling einer großen Anzahl von schwach abhängigen oder unabhängigen Aufträgen
  - „Utilizing otherwise idle workstations"
- On Demand
  - Kurzfristige Nutzung von Ressourcen, die lokal nicht kosteneffektiv bereitgestellt werden können
- Data Intensive
  - Zugriff auf weit verteilte umfangreiche Datenbestände
- Collaborative
  - Neue Formen der verteilten Zusammarbeit im Grid
  - Teleimmersion

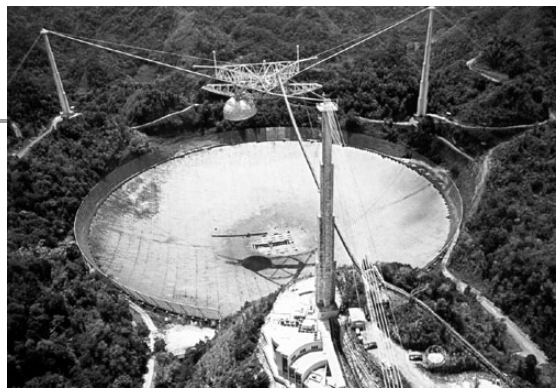
19.23

Beispiel: Leistungssteigerung

- Nutzung des Leistungspotentials „Internet"
- Bestimmte Anwendungsklassen
  - Distributed Supercomputing
    - Günstiges Verhältnis Rechnen-zu-Kommunikation
    - Beispiele: Mehrgittersimulationen, Ray-Tracing, …
  - High Throughput
    - Viele unabhängige Rechenaufträge
- Middleware-Ansätze
  - Ausgangspunkte PVM und MPI
  - Neuere Projekte: JXTA, Globus, Legion, …

19.24

Das Paradebeispiel für High Throughput seti@home



- Arecibo Observatorium
  - Puerto Rico
  - Durchmesser 305 m
  - Fläche 70000 m^2
- Radiodurchmusterung des Himmels
- Aufnahme des Radiosignals
  - 21 cm Wasserstoffband, Bandbreite 2 MHz
  - Aufteilung in 50 sec Abschnitte, Bandbreite 20 kHz
  - Suche nach künstlichen Signalen: Fourieranalyse
- 257 GByte Daten pro Woche
  - ca. 400 CD-ROM

19.26

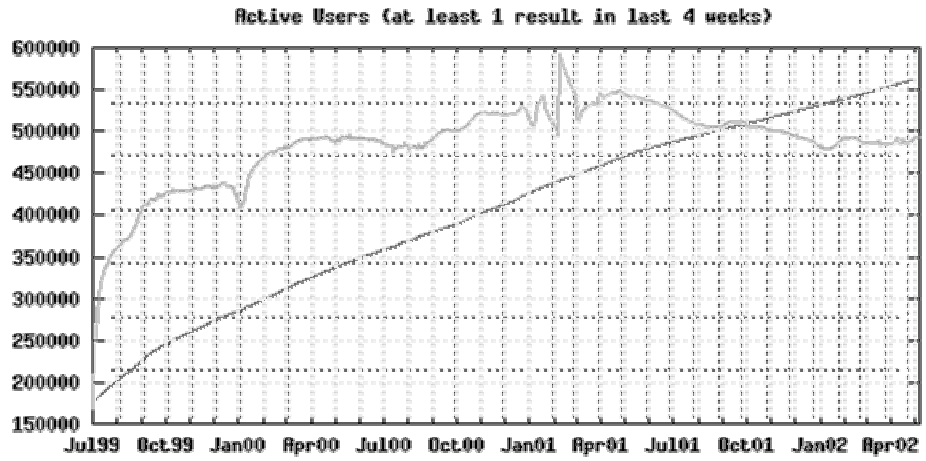
Client-Seite

- SETI@home Bildschirmschoner
  - Lädt Portionen auf den lokalen Rechner
  - Fourieranalyse während der Idle-Phasen
  - Übermittlung der Ergebnisse
- Aktuelle Zahlen (8.5.2002)

	Total	Last 24 Hours
Users	3711179	2604
Results received	504811241	1212902
Total CPU time	974194.727 years	1903.617 years
Floating Point Operations	1.552949e+21	4.730318e+18 (54.75 TeraFLOPs/sec)
Average CPU time per work unit	16 hr 54 min 18.7 sec	13 hr 44 min 54.8 sec

19.27

Weitere Daten



19.28

Germany

Last updated: Wed May 8 04:07:22 2002 UTC

The second annual German SETI@home meeting is scheduled for May 17-20 2002 in Weimar/Thüringen.

Name (and URL)	Results received	Total CPU time	Average CPU time per work unit	Date Last Result Received
1) Carlson	253330	248.445 years	8 hr 35 min 27.8 sec	Wed May 8 04:00:41
2) WelleErdball	211307	211.096 years	8 hr 45 min 04.5 sec	Tue May 7 17:49:29
3) Arnd	193363	213.216 years	9 hr 39 min 33.8 sec	Wed May 8 03:20:46
4) Billy the Mountain	164414	162.934 years	8 hr 40 min 52.1 sec	Wed May 8 04:04:42
5) SaschaH[SG1]	155567	137.371 years	7 hr 44 min 07.4 sec	Tue May 7 16:33:44
6) tiny island	150672	216.890 years	12 hr 36 min 35.6 sec	Tue May 7 08:03:36
7) (k.T.)	150312	157.033 years	9 hr 09 min 06.0 sec	Sun Apr 28 18:48:05
8) Hasso	148794	167.359 years	9 hr 51 min 10.6 sec	Wed May 8 03:53:31
9) birnsys	130770	122.981 years	8 hr 14 min 17.6 sec	Wed May 8 03:54:31
10) MagicVillage Team	121181	136.509 years	9 hr 52 min 05.0 sec	Wed May 8 01:11:29
11) Stephan Piel	113638	239.520 years	18 hr 27 min 49.7 sec	Tue May 7 22:45:32
12) SETI@uni-trier.de (I)	112994	186.626 years	14 hr 28 min 06.2 sec	Wed May 8 03:44:57

Diskussion

- Simple und traditionelle Client/Server-Architektur
  - Einfache Reaktion auf Fehlerfälle
  - Mehrere Stunden Rechenzeit zu wenige Sekunden Kommunikationszeit
  - Keine Nutzung spezieller Grid-Middleware ($\Rightarrow$ einfach installierbar)
- … aber mit durchschlagendem Erfolg
- Nachahmer in vielen Bereichen
  - Verteiltes Knacken von großen Schlüsseln
  - …

19.30

Allgemeine Grid Middleware

- Eigener Sprachansatz
  - Mentat (Parallele Fassung von C++)
  - Betonung reflektiver Mechanismen (Metaklassen)
- Beschreibung von Schnittstellen
  - Legion-aware Implementierungen
  - Wrapping von anderen Komponenten (MPI, PVM, …)
- Ganzheitlicher Ansatz
  - GRID-Betriebssystem

19.32

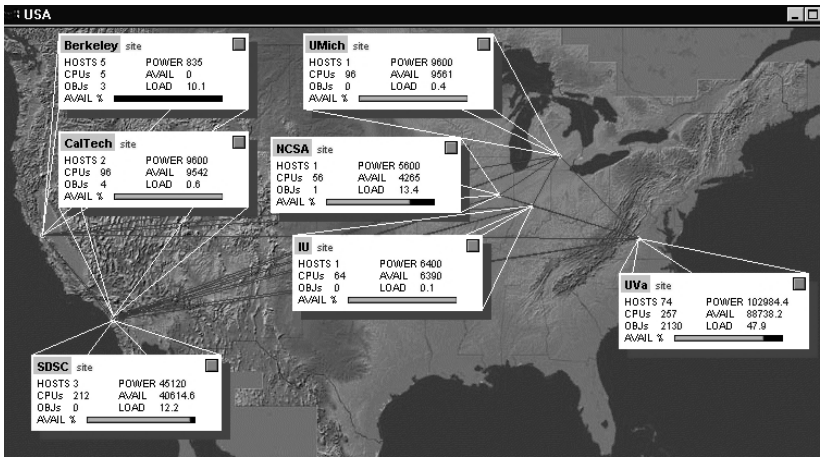
Tools

- MPI / PVM
- P-space studies - multi-run
- Parallel C++
- Parallel object-based Fortran
- CORBA binding
- Object migration
- Accounting
- Remote builds and compilations
- Fault-tolerant MPI libraries
- Post-mortem debugger
- Console objects
- Parallel 2D file objects
- Collections
- Licence support

www.cs.virginia.edu/~legion

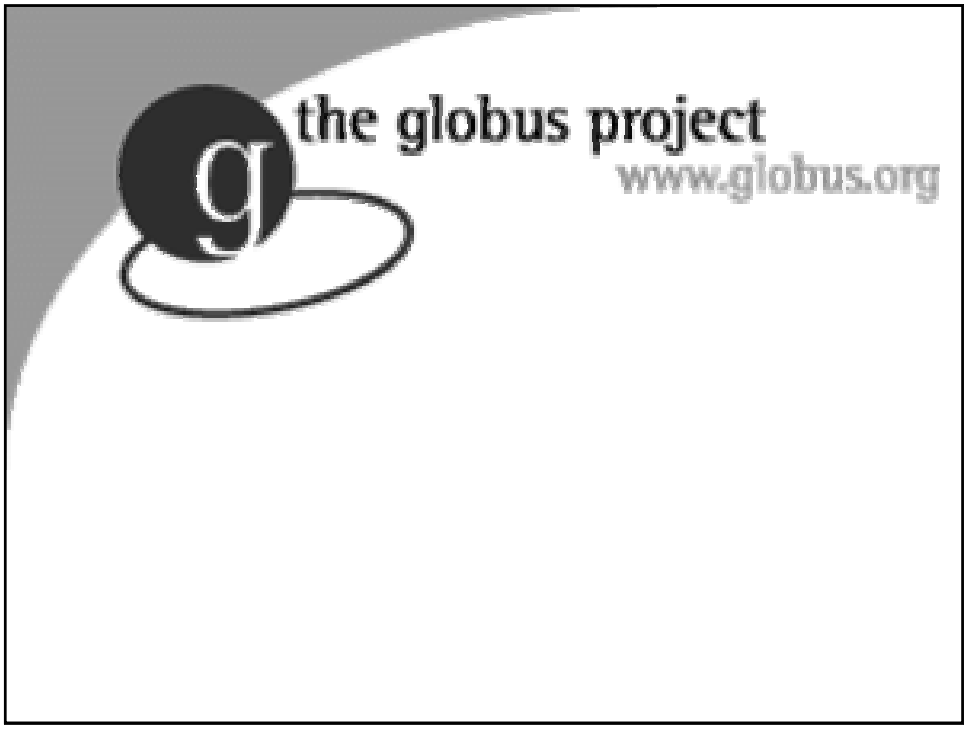
19.33

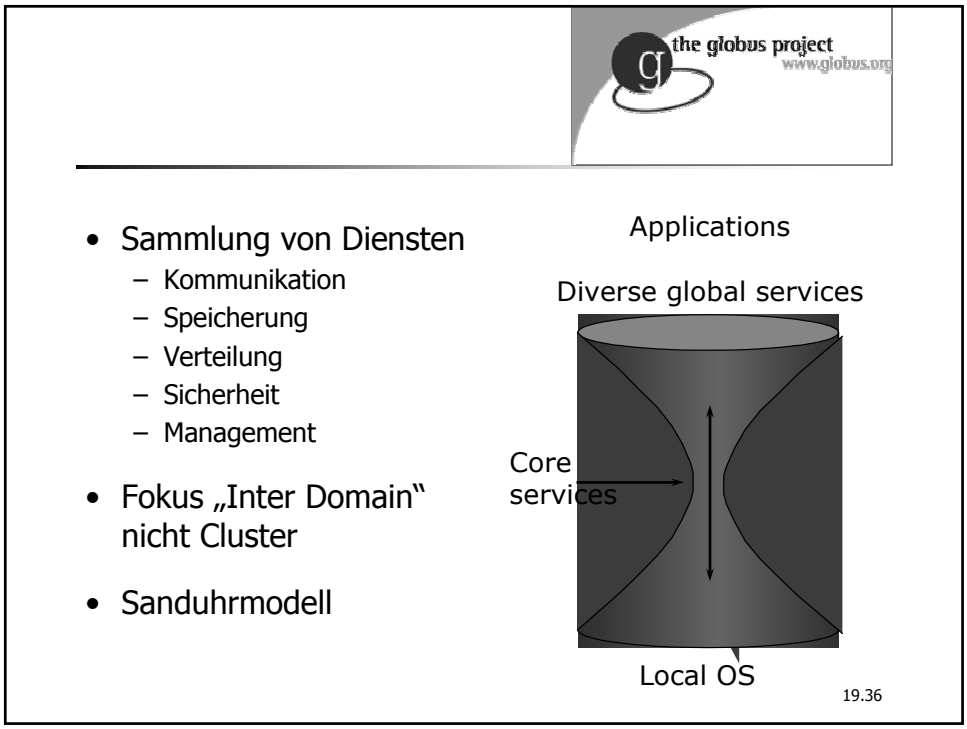
Beobachten



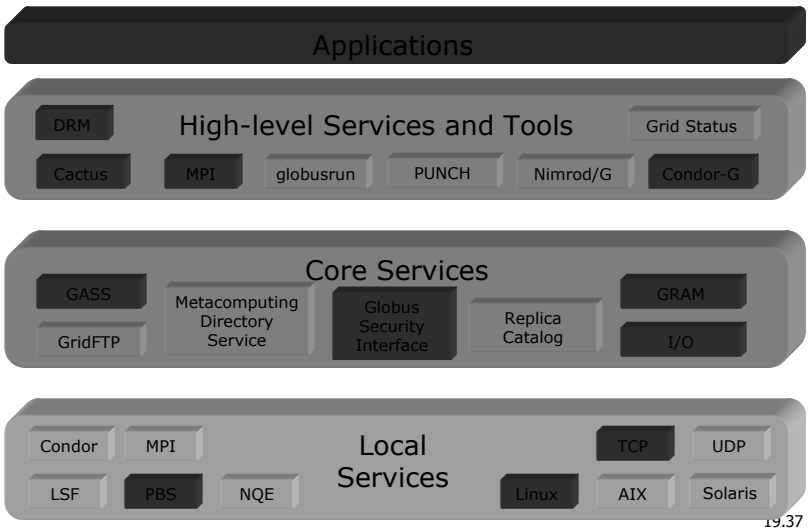
www.cs.virginia.edu/~legion

19.34


the globus project
www.globus.org

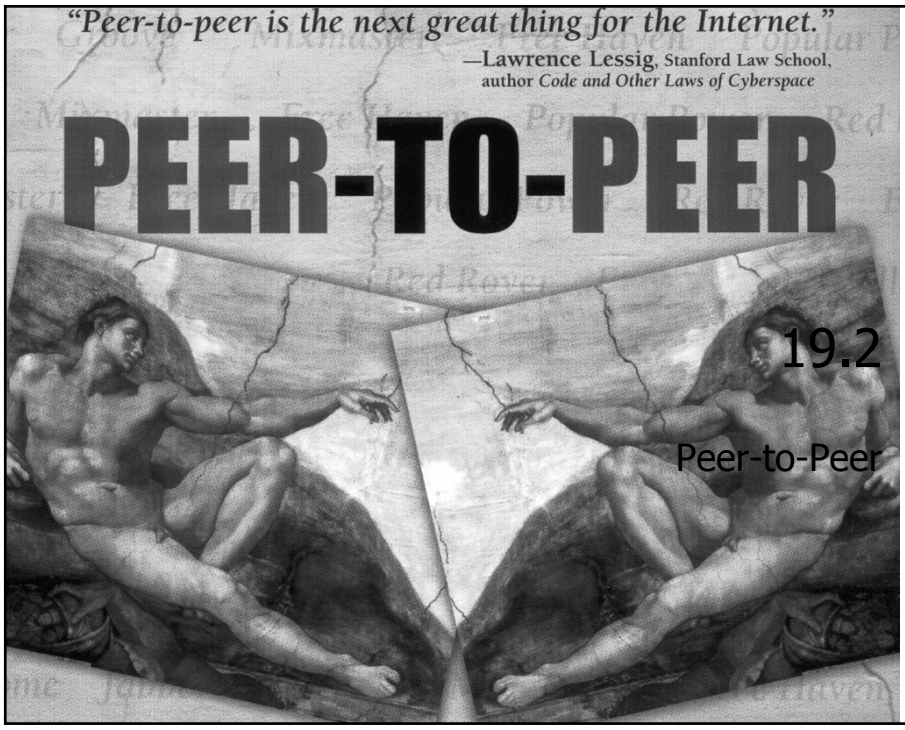

the globus project
www.globus.org
• Sammlung von Diensten
– Kommunikation
– Speicherung
– Verteilung
– Sicherheit
– Management
• Fokus „Inter Domain" nicht Cluster
• Sanduhrmodell
Applications
Diverse global services
Core services
Local OS
19.36

Architektur



19.37

aus www.globus.org



Titelbild des gleichnamigen Buchs aus dem O'Reilly-Verlag, Hrsg. Andy Oram, 2001

Zu P2P zählt ...

- Verteiltes Rechnen, Grid-Computing
- Austausch von Daten / Dokumenten
  - Forschung: P2P-Dateisysteme
  - Beispiele: Napster, Gnutella, Freenet
- „Freies Internet"
  - Anonymität und keine Rückverfolgung
    - bei Mail (Remailer)
    - bei Dokumenten (Free Haven)
  - Keine Zensur und keine Fälschungen im Web (Publius)

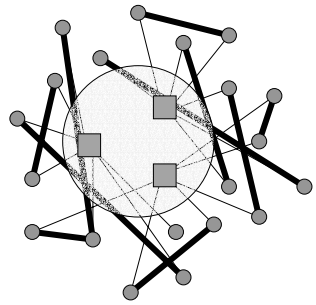
19.39

Austausch von Daten

- Verlagerung des Verkehrs in den Internet-Rand
  - Bessere Gesamtauslastung
  - Jeder Teilnehmer ist Server und Client
- Ausgangspunkt MP3-Tauschbörsen
  - Napster (Niedergang wegen rechtlicher Probleme)
  - Aktuelle Alternativen: Gnutella-Netze mit diversen Clients
- Verteilte Dateisysteme
  - CFS (MIT) Peer-to-Peer Read-Only Cooperative Storage

19.40

Beispiel Napster



- Trennung in
  - Daten
    - MP3-Musikstücke
  - Metadaten
    - Interpret, Titel, Adreßinformation, …
- Speicherung der Metadaten auf zentralen Servern
- Client
  - Suchanfrage an einen zentralen Server
  - MP3-Datei(en) von anderen Peers laden
- Zentrale Speicherung der Metadaten verletzt Schutzrechte

19.41

- Keine zentrale Speicherung der Metadaten
- Beeindruckendes Datenvolumen zugreifbar (z.B. ca. 5 Minuten nach Start eines Client)
  - 9000 bekannte Hosts
  - 19 Milliarden tauschbare Dateien
  - 64 TByte Gesamtgröße
- Konsequenzen
  - Hohe Netzlast über weite Bereiche
  - Teilweise fragwürdige Inhalte ladbar
  - Sehr geringe Zuverlässigkeit

19.42

Gnutella-Netze

- Rechner sucht und hält Verbindung zu k Peers
  - Host Lists, Host Caches, Verkehr beobachten, Mail, …
  - TCP-Verbindungen (PINGs)
- Initiator
  - Sendet Query an aktuell verbundene Peers
- Peer
  - Weiterleiten an eigene Peers (Flooding des Overlay-Netzes)
  - Falls Treffer, Rückmeldung an Initiator
- Qualität der Peers bestimmt den Erfolg
  - Small World Model, Power Law
  - Manche Peers sind „gleicher"

19.43

Sicherheits- und Schutzaspekte

- Grundgedanke: Gleichheit zwischen Peers
  - Aber überall gibt es Falschspieler
  - Und eingeschränkte Überwachbarkeit durch Staat etc.
- Schutz gegen Falschspieler
  - Aufbau von Vertrauen (Trust)
  - Verantwortung (Accountability)
  - Reputation
- Eingeschränkte Überwachbarkeit
  - Keine Zensurmöglichkeit
  - Anonymität

19.44

Analogie zu mobilen Systemen

- Ähnlichkeiten zwischen
  - Stationäre Peers mit ständig wechselnden Adressen
  - Identifizierbare Einheiten mit ständig wechselndem Standort
- Lösungsansätze in einer Forschungsrichtung häufig übertragbar
- Mobile Systeme (Ubiquitous Computing)
  - Keine stationäre und zuverlässige Kommunikationsbasis
  + Kein Pseudospoofing durch physische Gegenwart

19.45

19.3

Einfache und massive Multiplayer-Spiele

The END

- Ernüchterung nach Euphorie der ersten Jahre
  - Leistungssteigerung: Meist nur durch erneutes Programmieren
  - Fehlertoleranz: Enorme Materialschlacht
  - Natürliche Verteilung: Mensch denkt eher sequentiell
- Inhärente Probleme
  - Keine globale Zeit
  - Kein globaler Zustand
  - Indeterminismus
  - Ausfälle
  - ...
- Reizvoll
- Verteilte Systeme trotz aller Widrigkeiten extrem wichtig
  - WWW, E-Commerce, Multiplayer, ...

19.48

Literatur

- I. Foster, C. Kesselman (Hrsg.)
  The GRID - Blueprint for a New Computing Infrastructure
  Morgan Kaufmann, 1999
- A. Geist, A. Beguelin, J. Dongarra, W. Jiang, R. Manchek, V. Sunderam
  PVM 3 User's Guide and Reference Manual
  ORNL/TM-12187, 1993
  typischerweise Teil einer PVM-Distribution
- G.F. Pfister
  In Search of Clusters
  2nd Edition, Prentice-Hall, 1998

19.49