



HPC-Benutzerkolloquium

G. Hager/T. Zeiser

25.01.2005
RRZE

Agenda



- **EM64T-Clustererweiterung**
 - Allgemeines (GH)
 - Zukünftige Konfiguration: Queues, Policies, Betriebsmodus (GH)
 - Verfügbare Compiler und MPI-Versionen (TZ)
 - Modules-System (TZ)
 - Experimentelle Software (TZ)
- **Beschaffungen am LRZ**
- **Status Nachfolgebeschaffung VPP300 am RRZE**
- **"Upcoming Events"**

25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

2



EM64T-Clustererweiterung am RRZE



EM64T-Clustererweiterung



- **SFB 473 der DFG:**
Mechanisms of Transcriptional Regulation
- **HBFG-Antrag gestellt am 21.06.2004**
 - **Volumen: 240000 €**, verdoppelt aus
 - 100000 € von FAU
 - 20000 € aus KONWIHR/FH Nürnberg
 - **DFG Eingang am 09.07.2004**
 - **positiv begutachtet am 13.10.2004**
 - **SFB beauftragt RRZE mit Vertragsverhandlungen, Beschaffung und Betrieb der Clustererweiterung**
 - **Vertragsunterzeichnung mit Transtec am 15.11.2004**
- **Erweiterung wurde ausdrücklich mit dem Ziel beantragt und beschafft, eng in das bestehende IA32-Cluster integriert zu werden**

25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

4

EM64T-Clustererweiterung



- **Anlieferung und Aufbau der Knoten am 14.12.2004**
 - 64 Rechenknoten + 1 Frontend + 1 Fileserver
 - Knoten: Dual Intel Xeon "Nocona" (EM64T), 3.2 GHz, 1 MB L3, 2 GB RAM, GBit Interconnect
 - **16 Knoten** zusätzlich ausgestattet mit PCI-X Infiniband-Karten plus 24-Port Switch
 - Fileserver: 3,2 TByte brutto, **2,4 Tbyte netto-Kapazität**
- **Beginn der Abnahmephase am 15.12.2004 12:00**
- **Ende der Abnahme: Mitte Januar 2005**
- **Während der Abnahme**
 - Integration in RRZE-Umgebung
 - Ausräumen von 32-/64-Bit-Problemen
 - Einrichten des Queueing-Systems

25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

5

EM64T-Clustererweiterung



- **Was ist neu bei Nocona/EM64T?**
 - 64-Bit Linux-Betriebssystem (analog Opteron)
 - Fähigkeit, 32-Bit und 64-Bit Programme nativ auszuführen
 - Im 32-Bit-Modus: Vorteile durch
 - verdoppelten Cache
 - höhere Taktfrequenz
 - SSE3 und weitere architekturelle Verbesserungen
 - Im 64-Bit-Modus: Weitere Vorteile durch
 - breitere Integer-Register (64 statt 32 Bit)
 - doppelte Anzahl SSE2-Register
- **32-Bit-Compiler funktionieren weiterhin wie gewohnt**
 - **-xP** Flag nutzt SSE3-Erweiterung
- **Native 64-Bit-Compiler verfügbar**
 - damit auch neue MPI-Bibliotheken

25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

6

EM64T-Clustererweiterung



- **Was bringt Infiniband?**
 - PingPong-Werte:

	GBit Ethernet	Infiniband
Internode-Bandbreite	50-80 MB/s	600-800 MB/s
Internode-Latenz	25 μ s	5 μs

- **RRZE-Installation nutzt PCI-X Karten**
 - Bandbreite begrenzt auf 1 GB/s absolut
- **IB-Libraries und MPI gibt es nur für 64 Bit**
- **Aktuell: Testbetrieb!**
 - bei Interesse bitte melden

25.01.05

hpc@rrze.uni-erlangen.de

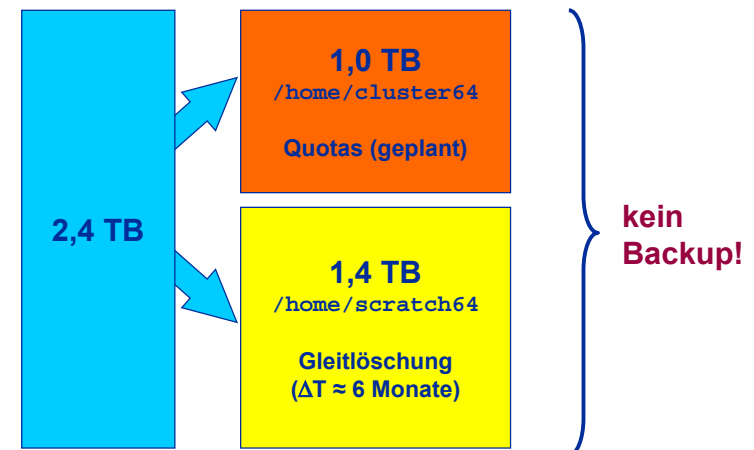
HPC-Kolloquium

7

Cluster-Erweiterung: Plattensysteme



- **2,4 TByte Nettokapazität unter RAID5 + 1 Hotspare**



25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

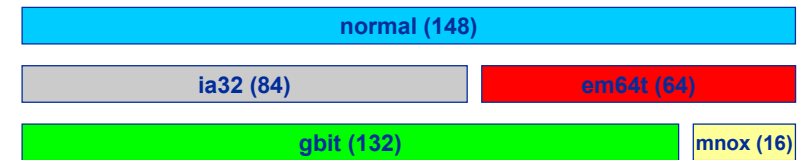
8

Cluster-Erweiterung: Partitionen



- Weiteres Frontend **sfront03** mit neuer Architektur
 - zur Kompilation und für interaktive Tests
 - TotalView-Debugger braucht erst noch Update
- Queueing-System muss so konfiguriert werden, dass
 - Unterschiede in der Knotenarchitektur adressierbar sind
 - bestimmte Vorgaben erfüllt werden können
 - trotzdem hohe Auslastung erreicht wird
- "Normale" Userjobs (Kurzläufer) vom alten Cluster sollen die Möglichkeit haben, freistehende Ressourcen der Erweiterung zu nutzen
- **Lösung: Aufteilung des kompletten Clusters in "Partitionen", die durch Queues oder andere Mittel adressierbar sind**

Cluster-Erweiterung: Partitionen



"node properties"

- "normal": alle Knoten
- "gbit": alle Knoten, die nur GE-Interconnect haben
- "ia32": alle "alten" Knoten
- "em64t": alle "neuen" Knoten
- "mnox": alle Knoten mit IB-Interconnect

2 weitere Knoten laufen unter Windows 2000 Server

Cluster: Batchbetrieb



- Queue-Konfiguration: runderneuert!

Queue	Laufzeit [HH:MM:SS]	min-max. CPUs/Job	wer darf?
express	≤ 01:00:00	1-8	alle
iexpress	≤ 01:00:00	1-8	alle
s1	01:00:01 ≤ T ≤ 06:00:00	1-64	alle
s2	06:00:01 ≤ T ≤ 48:00:00	1-64	alle
s3	48:00:01 ≤ T ≤ 168:00:00	1-8	auf Antrag
ls_normal	T ≤ 24:00:00	1-64	SFB
ls_long	24:00:01 ≤ T ≤ 240:00:00	1-8	SFB
iband	T ≤ 64:00:00	4-32	auf Antrag

- Weitere Queues (special, fhg, lsm) sind für Spezialzwecke vorgesehen

Cluster: Batchbetrieb



- Aufteilung der Queues auf die Knoten, weitere Beschränkungen

Queue	Knoten	max. RUNNING CPUs
express	ia32 (werktags 7-20h sind 4 Knoten reserviert)	alle (168)
iexpress	mnox (werktags 10-16h sind 4 Knoten reserviert)	alle (32)
s1	gbit	alle (264)
s2	ia32	alle (168)
s3	ia32	8
ls_normal	gbit	alle (264)
ls_long	gbit	64
iband	mnox	alle (32)

Cluster: Batchbetrieb



- Wer merkt sich das alles?
- Niemand! Einsortierung in die Standard-Queues erfolgt automatisch anhand
 - Laufzeit
 - Knotenzahl
 - Zugriffsbeschränkungen
- Normalbenutzer muss keine Queue mehr angeben, Default-Queue ist "route"
- "route" verteilt Jobs anhand der Ressourcen etc. auf andere Queues
- Ausnahme: Queue "iband" und andere Spezialqueues

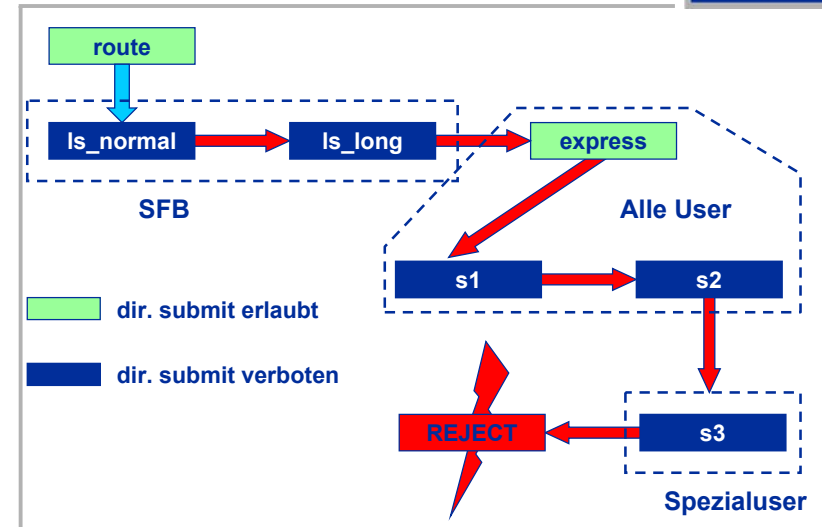
25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

13

Cluster: Batchbetrieb



25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

14

Cluster: Batchbetrieb



- Frage: Wie spezifizieren User in s1 oder ls_*, dass nur EM64T bzw. IA32 erwünscht ist?
- Angabe von "node property" beim Submittieren!

Beispiel: Langer SFB-Job, aber nur für EM64T (Default wäre alle GBit-Knoten):

```
qsub -l walltime=200:00:00,nodes=4:ppn=2:em64t ...
```

Beispiel: Kurzläufer eines "Normalusers", der nur IA32 will:

```
qsub -l walltime=04:00:00,nodes=2:ppn=2:ia32 ...
```

- Vorsicht! Kombination unmöglicher Properties führt dazu, dass Job nicht gescheduled werden kann:

```
qsub -l walltime=12:00:00,nodes=1:ppn=2:mnox ...
```



25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

15

Cluster: Batchbetrieb



- Resultat: Für "normale" User wird das Leben einfacher
 - Laufzeiten bis 48h für Jobs bis 64 CPUs
 - Vorteile für kurze Jobs (können neue Knoten mit benutzen)
 - Restartfähigkeit prüfen!
 - Queue-Angabe kann i.A. entfallen
 - direktes Submittieren in die s?-Queues ist nicht erlaubt
 - bei Bedarf nach langen Laufzeiten: in s3-Queue eintragen lassen
- Spezialuser SFB
 - landen per Default in ls_*-Queues (kein direktes Submittieren)
 - durch Angabe der "node property" Anforderung von EM64T
- Infiniband-Knoten
 - zugänglich auf Antrag über Queue "iband" bzw. "iexpress"
 - Aktuell im Spielbetrieb
 - Erweiterung auf 24 IB-Knoten geplant

25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

16

Cluster: Batchbetrieb



- Umstellung auf die neue Queue-Struktur
 - "route"-Queue ist bereits aktiv und kann benutzt werden
 - Sperrung der "alten" Queues

serial, parallel, serial_long, bco_normal, bco_long, hmh

im Laufe der kommenden Tage (nach Ankündigung), d.h. Queues laufen noch leer, Submit ist aber nicht mehr möglich
 - gleichzeitig: "route" als Default-Queue (statt express)
 - gleichzeitig: Doku-Update (Webseite)
- Diskussion

25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

17

Beschaffungen am LRZ: HLRBII und VPP-Nachfolge



Beschaffung HLRBII am LRZ



- Ausschreibung fand unter Mitwirkung des RRZE statt
 - Benchmarks, Bewertung, Mitarbeit im Auswahlgremium
 - FAU-Benchmarks: TRATS (LSTM), DMRG (RRZE), SIPBench (RRZE), LASER (LSS)
 - Volumen: **38 Mio. € + Betriebskosten**
- Ende November 2004: Entscheidung für SGI-Angebot aufgrund der besten zugesagten Applikationsleistung
- Ergebnis: SGI Altix "Tornado" System
 - Numalink4 Interconnect (doppelt so schnell wie RRZE_System)
 - Dual-Core Itanium2 (Montecito)
 - Bandbreite pro CPU besser als RRZE-System
 - Im Regelbetrieb Partitionen von 1024 CPUs pro SSI

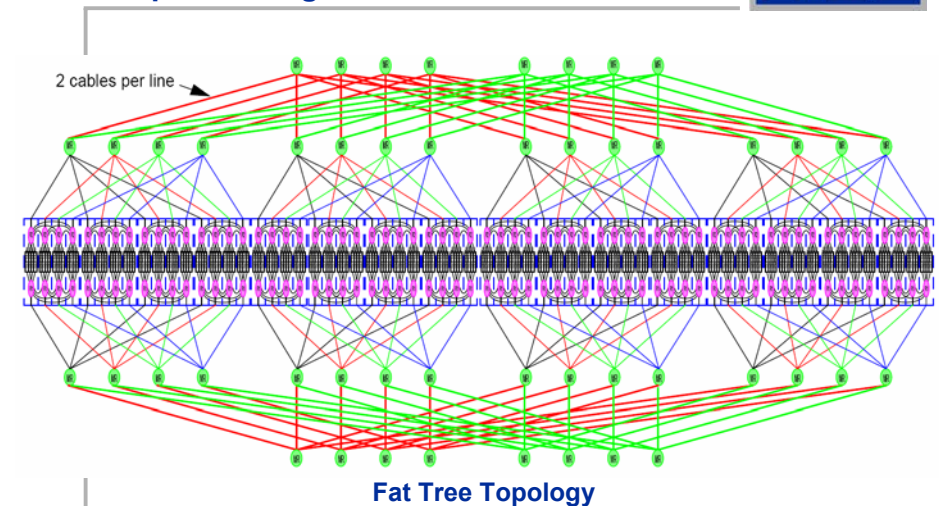
25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

19

SGI Tornado - 256 Processor Chips Building Block



25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

20

Die Software für die SGI Altix Systemfamilie



Standard SUSE Linux Betriebssystem

unterstützt Systeme zwischen 2 und 512/1024 Prozessoren (SSI) in einer einzigen Instanz des Betriebssystems

SGI Software (ProPack) für die besonderen Anforderungen großer HPC Systeme

MPI, SCSL, HISTX, PCP, XVM,...

Intel Compiler und Toolset

Unterstützung aller gängigen Programmiermodelle

MPI, OpenMP, Pthreads, auch gemischt unter Ausnutzung spezieller Hardware Funktionen

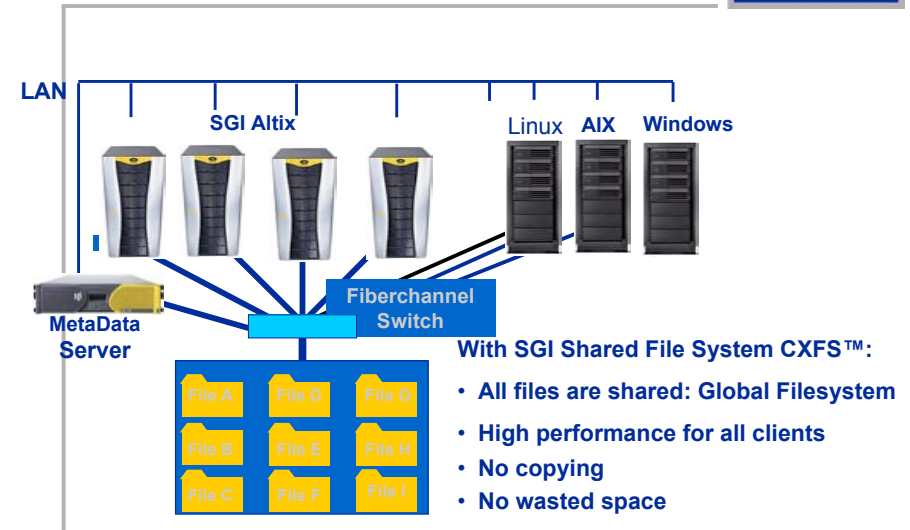
25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

21

Storage Area Network SGI® Shared File System CXFS™



One Global Filesystem

25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

22

HLRB II Zeitplan



Test- und Migrationssystem im Juli 2005

- 64 CPUs, 256 Gbyte Hauptspeicher, 6 Tbyte Plattenkapazität
- Portierung/Optimierung/Skalierung der Programme
- Installierung und Test der Betriebsumgebung

Phase 1 (März 2006)

- 5120 Montecito Prozessor Cores (2560 Prozessor Chips)
- 20 Tbyte Hauptspeicher
- CXFS Shared File System, Storage, I/O, Netzzugang

Phase 2 (Juni 2007)

- 6656 Montvale Prozessor Cores (3328 Prozessor Chips)
- 40 Tbyte Hauptspeicher
- CXFS Shared File System, Storage, I/O, Netzzugang

Aufrüstung ohne Betriebsunterbrechung

25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

23

HLRB II Zeitplan



	Phase 1 (2006)	Phase 2 (2007)
Prozessor Typ	dual-core	dual-core
Cache Größe	6 MB	9 MB
Peak Leistung	> 6 Gflop/s	> 10 Gflop/s
Anzahl Prozessor Chips	2560	3328
Anzahl Prozessor Cores	5120	6656
Peak Gesamtleistung	32 Tflop/s	69 Tflop/s
Gesamter Hauptspeicher	20 Tbyte	40 Tbyte

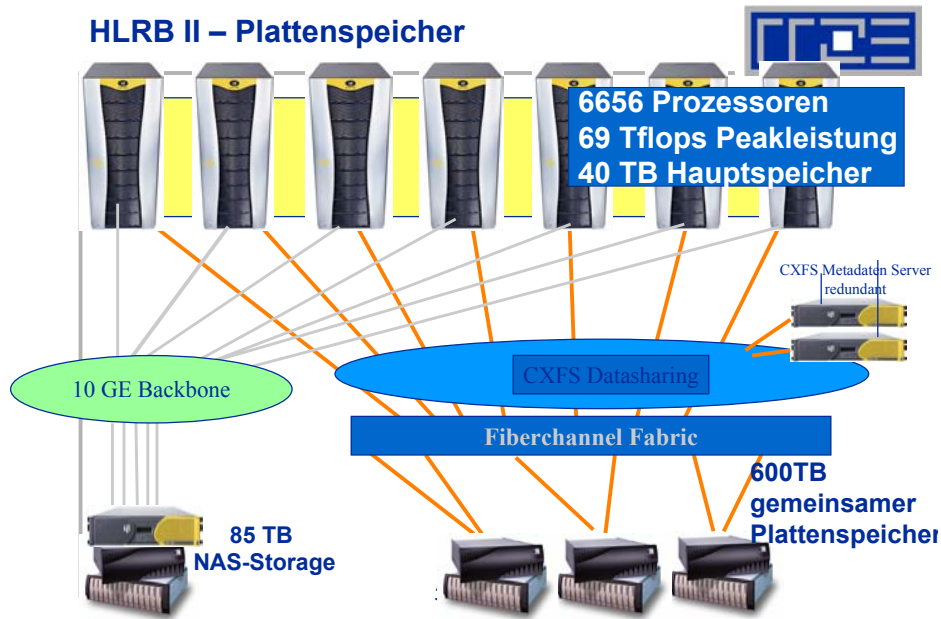
25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

24

HLRB II – Plattenspeicher



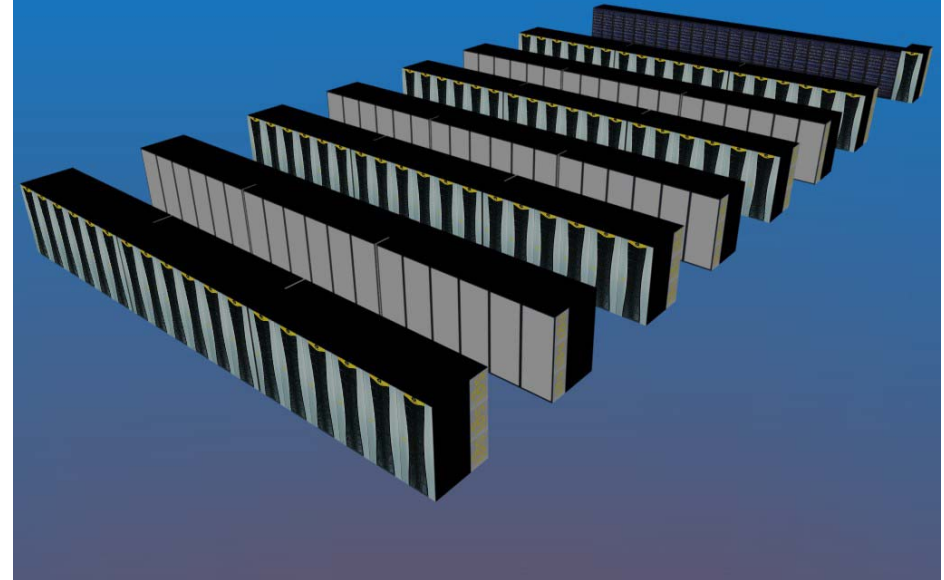
25.01.05

hpc@rrze.uni-erlangen.de

HPC-Kolloquium

25

Das HLRBII-System



VPP700-Nachfolge am LRZ



- Installation Februar 2005
- 128-CPU Altix "BX2"
- Madison 9M single-core CPUs mit 6MB L3 Cache (!) und 6.4 GByte Memory-Bandbreite pro Knoten
- doppelte Dichte (CPUs/Rack) wie RRZE-Altix
- NUMALink4
- de facto "Erweiterung" des LRZ-Linux-Clusters

25.01.05

hpc@rrze.uni-erlangen.de

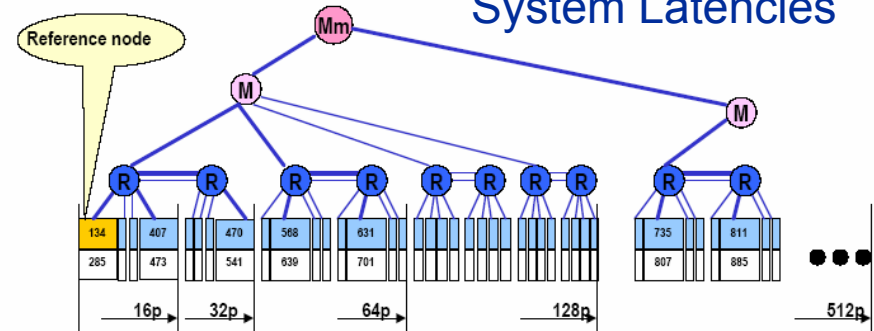
HPC-Kolloquium

27

ascender.americas: 512x1.3GHz/3MB - Jan 6, 2004

2p per brick

System Latencies



Local: 134ns **Approximate latency breakdown (ns)**

Remote = Local + 10ns + ... (10ns penalty for leaving node)

Xn = 70ns

Cable = 10 ns/m

First NL3 Router = 100ns (SHub-NL3 serialization)

Following NL3 Routers = 57ns

E.g. most remote:

Local + Remote + 3*Xn + 3*1m + 2*2m + 1*3m + 1*4m + FirstRouter + 5*FollowingRouter
134 + 10 + 3*70 + 3*10 + 2*20 + 30 + 40 + 100 + 5*57 = 873ns

sgi



- **10 Februar 2005, 9:00-17:00, LRZ (HALT1W04):**

"Einführung in die Nutzung der SGI Altix"
Vortragender: Rüdiger Wolff, SGI

Übertragung der Veranstaltung ans RRZE (per Videokonferenz) ist geplant, Anmeldung bitte an hpc@rrze.uni-erlangen.de

- **8./9. März 2005, RRZE:**

"Application Performance Optimization"
Fallstudien anhand realer Anwendungen, Profiling Tools, eventuell vertiefender Vortrag zum Thema Itanium2 Profiling