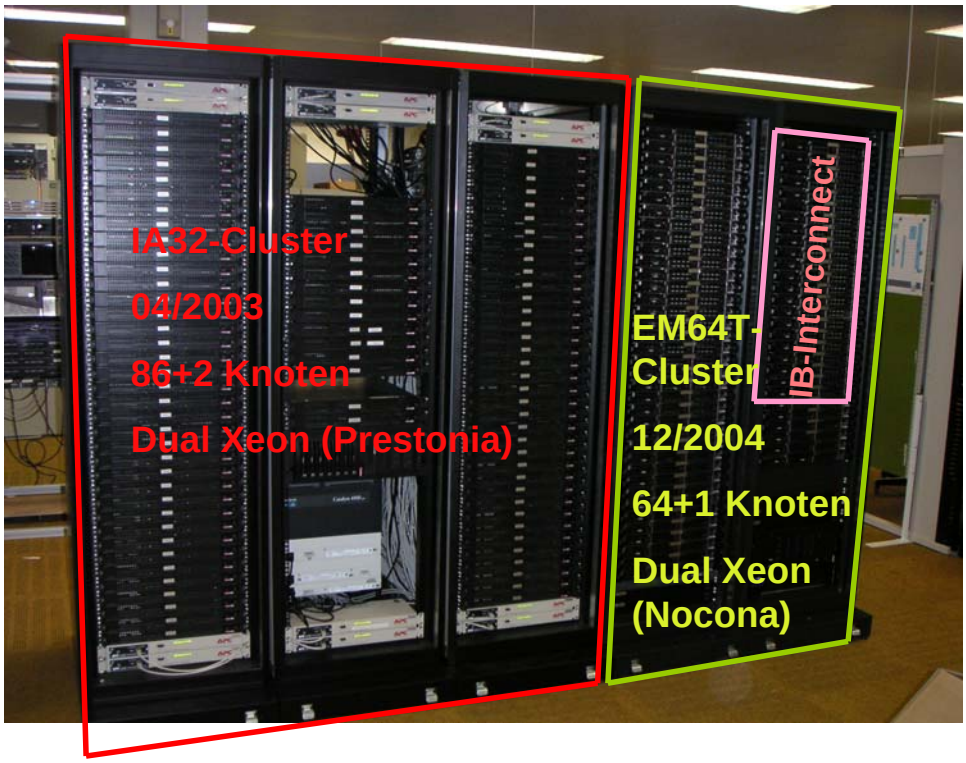
Betrieb eines heterogenen Clusters

Georg Hager
Regionales Rechenzentrum Erlangen (RRZE)

ZKI AK Supercomputing
Karlsruhe, 22./23.09.2005

Transtec GBit/IB-Cluster am RRZE



EM64T-Clustererweiterung

- SFB 473 der DFG:
  Mechanisms of Transcriptional Regulation
- HBFG-Antrag gestellt am 21.06.2004
  - Volumen: 240000 € (davon 20000€ von 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

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 (jetzt 24) 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
- Während der Abnahme
  - Integration in RRZE-Umgebung
  - Ausräumen von 32-/64-Bit-Problemen
  - Einrichten des Batch-Systems

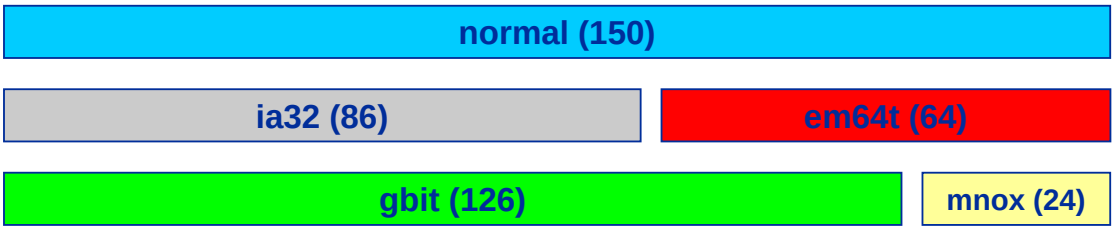
EM64T-Clustererweiterung

- 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 Architekturverbesserungen
  - 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

Clustererweiterung: Vorgaben

- Gemeinsames Batchsystem (OpenPBS+Maui)
  - neue Knoten werden bevorzugt an SFB vergeben
  - hohe Auslastung wird sichergestellt
    - bei Leerständen dürfen "normale" User die neuen Knoten benutzen
  - Generell kann ein bestimmter Knotentyp angefordert werden (IA32, EM64T, Infiniband)
- Entwicklungsumgebung
  - Konsistente Compilerinstallation auf beiden Clusterteilen
  - IA32-Binaries laufen ohne Änderung auf den EM64T-Knoten (incl. MPI)
  - EM64T-Binaries können nur auf EM64T-Frontend erzeugt werden
  - TotalView (leider) nur für IA32 ($)

Clustererweiterung: 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

Cluster: Batchbetrieb

- Queue-Konfiguration

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, lstm) sind für Spezialzwecke vorgesehen

Cluster: Batchbetrieb

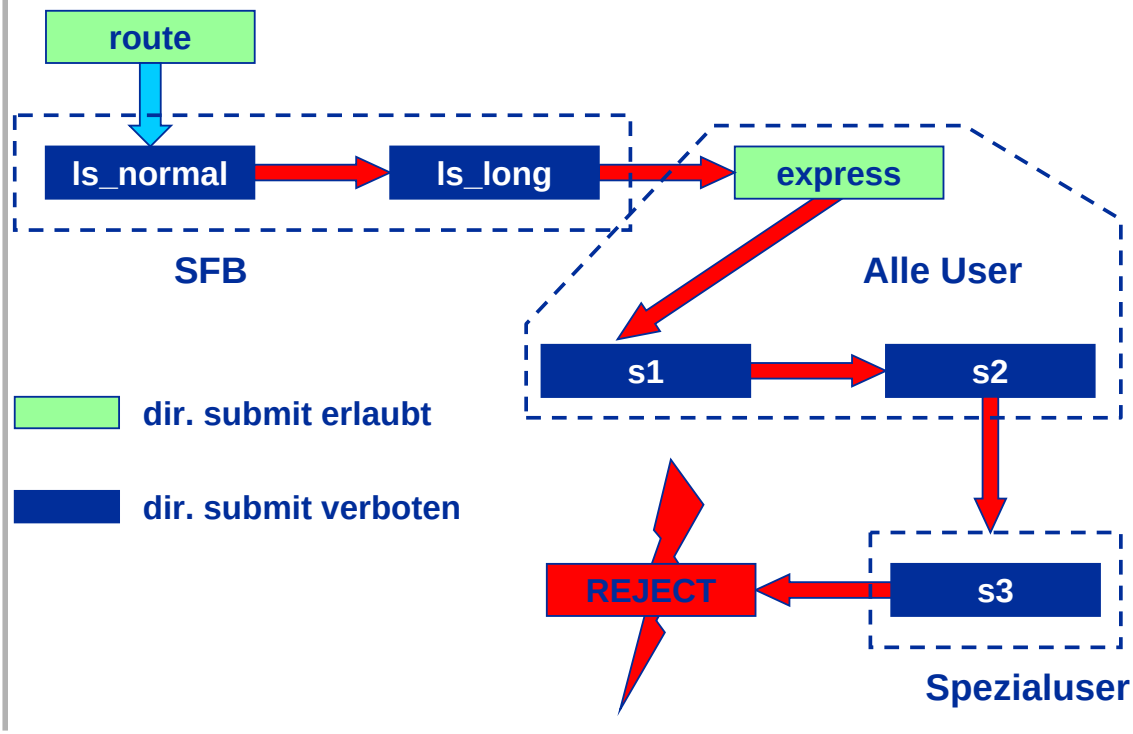
- Aufteilung der Queues auf die Knoten, weitere Beschränkungen

Queue	Knoten	max. RUNNING CPUs
express	ia32 (werktags 8-21h sind 4 Knoten reserviert)	alle (172)
iexpress	mnox (werktags 10-16h sind 4 Knoten reserviert)	alle (48)
s1	normal	alle (300)
s2	ia32	alle (172)
s3	ia32	8
ls_normal	gbit	alle (96)
ls_long	gbit	80
iband	mnox	alle (48)

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

Cluster: Batchbetrieb



Integration der neuen Knoten

- **Betriebssystem: Debian 3.0**
- **Hauptproblem: Integration der Intel-Compiler in die Debian 3.0 Entwicklungsumgebung im 64-Bit-Betrieb**
  - **Intel-Compiler benutzt Header und Bibliotheken der GNU Compiler Collection**
  - **Parallele Installation mehrerer GNU-Compiler plus Intel-Compiler ist problematisch**
  - **"Hacks" (Wrapperskripte um GNU-Compileraufrufe, Patches in Intel-Compiler-Konfigurationsfiles, GXX_INCLUDE) machen beide Welten kompatibel**
  - **Aktuelle GNU-Compiler werden vom RRZE unter separaten Pfaden zur Verfügung gestellt**
- **Bereitstellung geeigneter MPI-Bibliotheken erforderlich (32/64 Bit, compilerabhängig)**

Fazit: Wie läuft es nach 9 Monaten?

- Häufigste Userbeschwerden
  - "Mein Job läuft nicht los"
    - Maui gibt darauf üblicherweise eine erschöpfende Auskunft
    - statistisch bekommen SFB-User genau "ihren" Share an Zyklen (ca. 35-40%)
  - "Da sind Physiker auf unseren Knoten"
    - Kurzläufer sind überall erlaubt
  - gcc funktioniert nicht
    - Problem mit den gcc-Wrappern
  - "Ich brauche Paket X in Version 3.14, aber es ist nur in Version 2.72 vorhanden"/"Ich brauche eine 64-Bit-Version von Bibliothek Y"
    - Bereitstellung von (Back)ports, meist seitens Transtec
- Ansonsten: 90% durchschnittliche Auslastung!