EKPplus: Ein Linux-Cluster zur Auswertung von Hadron-Kollider Daten

Yves Kemp, Thomas Müller,
Günter Quast, Patrick Schemitz
Experimentelle Kernphysik, Univ. Karlsruhe (TH)

Motivation:

- Entwicklungsumgebung für Experiment-Software und individuellen Analysecode
- Echtzeitzugriff auf Computing-Ressourcen
- Schnelle Grid-Anbindung zum FZK
- Test auf Production-Level für Grid-Applikationen

Der Linux-Cluster EKPplus

Inhalt:

- Aufbau: Hardware
- Aufbau: Netzwerk
- Batch Queueing System OpenPBS
- Das Bootkonzept
- Performance: Datenbank-Limit
- Performance: Netzwerk-Limit
- Performance: CPU-Limit
- Ausblick / Ausbau

Hardware I

- 18 Rechenknoten:
  - Athlon 1300 MHz / Athlon XP 1600+
  - (512 MB oder 1GB) SD-RAM (DDR-RAM nicht lohnend)
  - 40 GB IDE-Platte → Zwischenspeicher
  - 100 Mbit Ethernet, bootfähig
- 2 Fileserver:
  - jeweils 2x 3ware Escalade 7850 → je 8 IDE-Platten
  - einmal 16x 100 GB IDE-Platten als RAID5 (1.4 TB netto)
  - einmal 16x 250 GB IDE-Platten als RAID5 (3.5 TB netto)
  - jeweils 2x Gigabit Ethernet
- Kontrollrechner

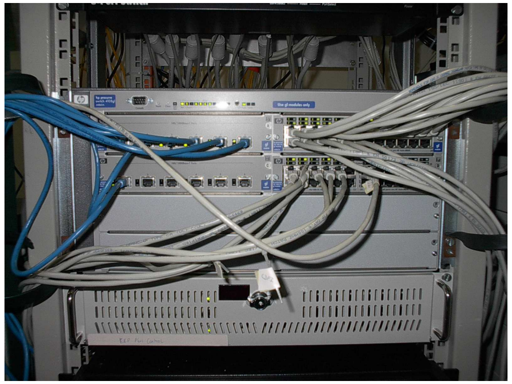
Hardware II

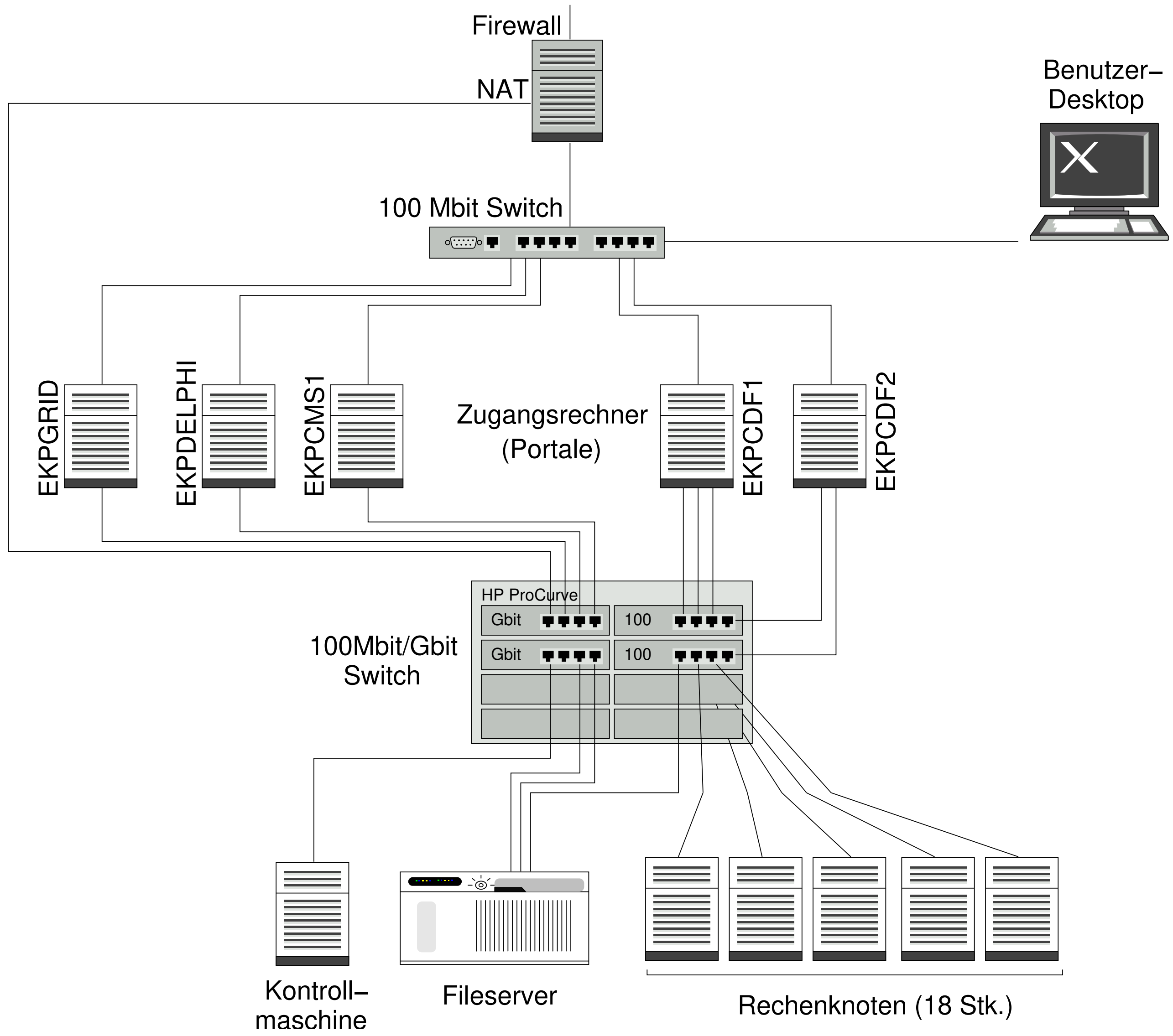
- 2 CDF-Portale:
  - jeweils Dual Athlon 1400 MHz bzw XP 1600+
  - jeweils 2 GB DDR-RAM
  - jeweils 4x 80 GB IDE-Platten als RAID0
  - jeweils 4x 100 Mbit Ethernet (Chipsatz: Probleme mit Gbit)
- CMS-Portal:
  - Dual Athlon MP 2100+
  - 2 GB DDR-RAM
  - 8x 120 GB Platten (4 als RAID0, 4 als RAID5)
  - Gigabit Ethernet (Intel)

Netzwerk - Hardware

HP ProCurve 4108gl

- 36 Gbit backplane
- 8 mögliche Erweiterungs-Einschübe
- 2 belegt durch Kupfer-Gbit Module (insgesamt 12 Ports)
- 2 belegt durch 100 MBit Module (insgesamt 48 Ports)




Firewall
NAT
Benutzer–
Desktop
100 Mbit Switch
EKPGRID
EKPDELPHI
EKPCMS1
Zugangsrechner
(Portale)
EKPCDF1
EKPCDF2
HP ProCurve
Gbit
100
Gbit
100
100Mbit/Gbit
Switch
Kontroll–
maschine
Fileserver
Rechenknoten (18 Stk.)

Netzwerk - logische Struktur

Routing → vier Subnetze ⇒ load-balancing:

- öffentliche IP-Adressen:
  Außenanbindung der Portale
- 192.168.101.0/24:
  Batchsystem; Knoten: Booten und System
- 192.168.102.0/24:
  NFS: Portal /home,
  Fileserver Vol. 1+3
- 192.168.103.0/24:
  NFS: Portal /scratch,
  Fileserver Vol. 2+4

Batch Queueing System

Verwendet wird OpenPBS.

- submittieren etc. von den Portalen
- vier Queues:
  short (1h), medium (8h), long (5d), IO (5d)
- bis zu 3 Jobs pro Knoten
- Job := Shellskript plus `#PBS`
- `http://www.openpbs.org`
- Rutherford, FZK Tier 1 Zentrum
- kostenlos
- sammeln Erfahrung mit Condor

Bootkonzept

Kontroll-Rechner:

- SuSE 8.0
- Standard-Bootvorgang
- NFS-Server für Knoten: `/bin, /lib, /usr, /sbin, /tftpboot/192.168.101.*`
- TFTP-Server für Bootloader `pxegrub` und Kernel

Knoten:

- DHCP $\rightarrow$ Netzwerkeinstellungen, TFTP-Server-Adresse
- TFTP lädt Bootloader
- Bootloader lädt Kernel
- Kernel mountet / vom Kontrollrechner
- Initskripte
  - mounten NFS-Laufwerke
  - starten YP, PBS

Performance: Datenbank-Limit

CDF verwendet zentrale Oracle-DB für Kalibration etc

Test: Dauer einer leeren Anfrage:

- Von einem Rechner der EKPplus aus: 8.8 s
- Von einem Rechner des FZK aus: 7.5 s
- Von einem Fermilab-Rechner aus: 0.3 s

Kein Unterschied Knoten/Portal

Konstant über den Tag

Großer Unterschied Fermilab/Übersee

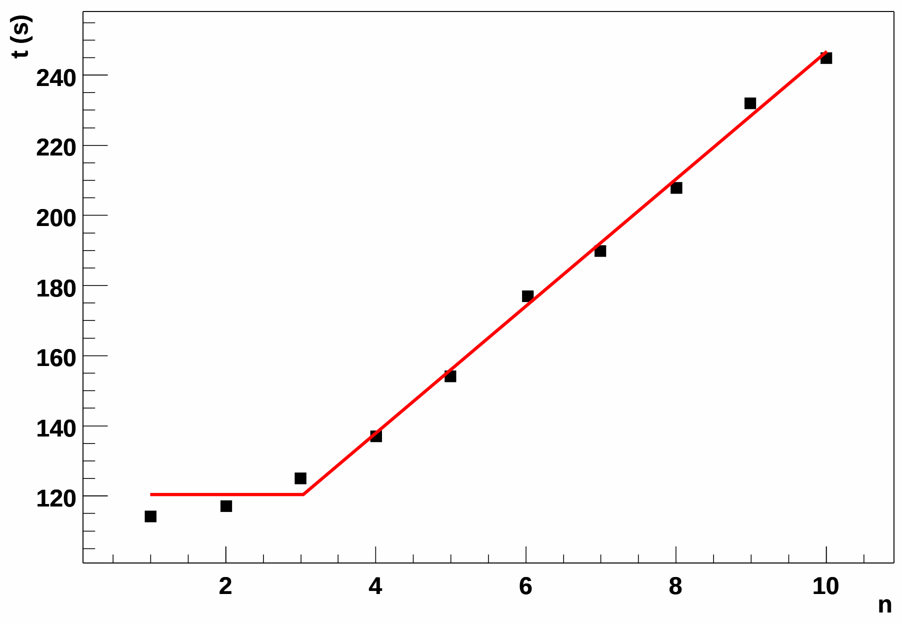
$\rightarrow$ Nachteil zentraler Datenbanken

Performance: Netzwerk-Limit

Test: Der gleiche Job läuft gleichzeitig auf n Knoten

Jobprofil: CDF-Analysejob mit hohen IO-Anteil, ROOT-File auf einem mit 100Mbit angebundenen NFS-Server

Zeit pro Job bei n Knoten:



Performance: CPU-Limit

- Produktion des “offiziellen” $t\bar{t}$-MC der CDF top-Gruppe
  29 s / Ereignis (Athlon 1300MHz).
  243 kB / Ereignis.

- Produktion von $b\bar{b}$ und $c\bar{c}$ mit cdfSim und anschliessend Rekonstruktion
  12 s / Ereignis (Athlon 1300 MHz)
  110 kB /Ereignis.

- Produktion von $H \to 4\mu$ mit Phytia/CMSSIM/Objectivity
  83 s / Ereignis (Athlon 1300 MHz)
  1.4 MB Plattenplatz /Ereignis.

Ausblick / Ausbau

- Plattenplatz: 6 TB (8 TB mit Austausch von Festplatten)
- Opteron-Testrechner?
- Ausgewählte Verbindungen auf Gbit hochrüsten
- Monitoring!!!