Linux an der GSI Darmstadt

1. Die GSI in Darmstadt
2. Idee und Realisierung der Linux Farm
3. Stand der Hardware
4. GSI Device
5. Software
6. Services und Monitoring
7. Batch Queueing System (LSF)
8. Zahlen
9. Probleme
10. Zukunft

1. GSI Darmstadt

Forschungsinstitut des Bundes und des Landes Hessen

Schwerpunkt Schwerionenforschung
- Suche nach neuen exotischen/schweren Kernen
- Plasmaphysik (Petawatt Laser Phelix)
- Materialforschung
- Biologie
- Kernchemie
- Medizin (Tumortherapie C12)
- Kern-/Atomphysik

DV/EE - Datenverarbeitung und Experimenteelektronik
- 3 große VMS Cluster
  - Beschleunigersteuerung (Unterstützung)
  - Rechenkapazität
- AIX
  - Massendatenspeicherung
  - Rechenkapazität
  - allgemeine Services
- Windows NT (2000)
- Linux
  - Rechenkapazität (LSF)
  - Desktop

2. Idee und Realisierung

- Aufbau eines zentralen Linux-Compute und Desktop Service
  (Ergänzung und Ablösung der existierenden Systeme)
  - Viele PC's (keine Obergrenze)
  - zentrale Server (File Server, Samba, Batch, Web, Printservice)
  - zentraler Dialog Service
  - Integration von dezentralen Desktop PC's
  - einfaches zentrales Management
  - logische Verknüpfungen innerhalb der Farm --> Art Cluster
  - (beliebige) Erweiterbarkeit
  - vergleichbare (identische) Hard- und Software
  - Batch Queueing System
- Ziel:
  - 4 Typen von Maschinen:
    - ohne X11 Unterstützung
      - Batch Server
      - Systemless Client (Batch)
    - mit X11 Unterstützung
      - Desktop Server
      - Systemless Client (Desktop)
  - Distribution muß Server/Client Konzept voll unterstützen

2. Idee und Realisierung

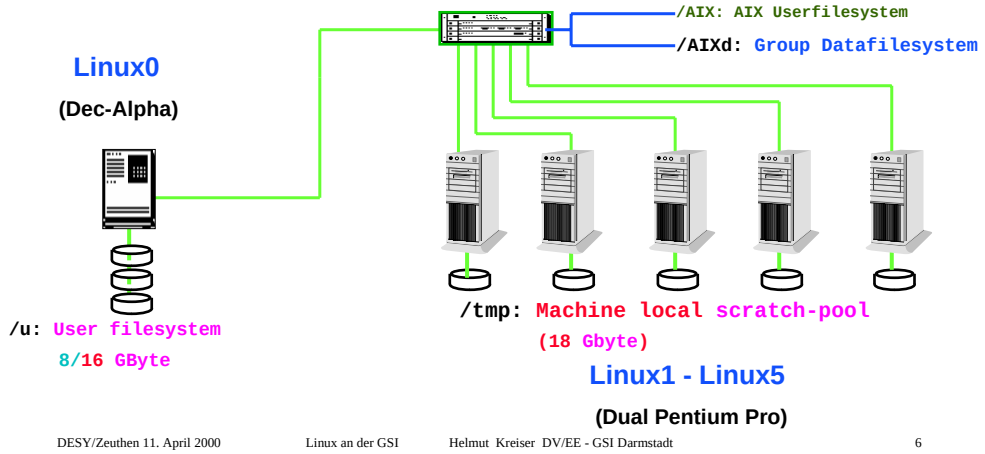
- Aufbau:
  - Betriebssystems-Server mit
    - Systemunabhängigen Teil (/usr)
    - Systemspezifischen Teil (/, /etc, /lib, /bin, /sbin)
      - Filesystem mit Pfaden für jeden Rechner
      - Booten eines zentralen, gemeinsamen Kernels
      - mounten des systemspezifischen Verzeichnisses vom Server (nfs-root)
  - Debian-Distribution gewährleistet Voraussetzungen (Stand vor 3 Jahren)
    - frei verfügbar im Netz, incl. Updates
    - große Entwicklergemeinde, keine Firma
      - aber kein Zugriff für spezielle Entwicklungen/Anpassungen möglich
    - Release-Zyklen sind relativ lang (stable, unstable)
    - wenig Änderungen (stable Release), evtl. (Security-) Patches
  - Keine Installationsprozeduren = Clonen (kopieren) der Basissysteme

2. GSI Rules

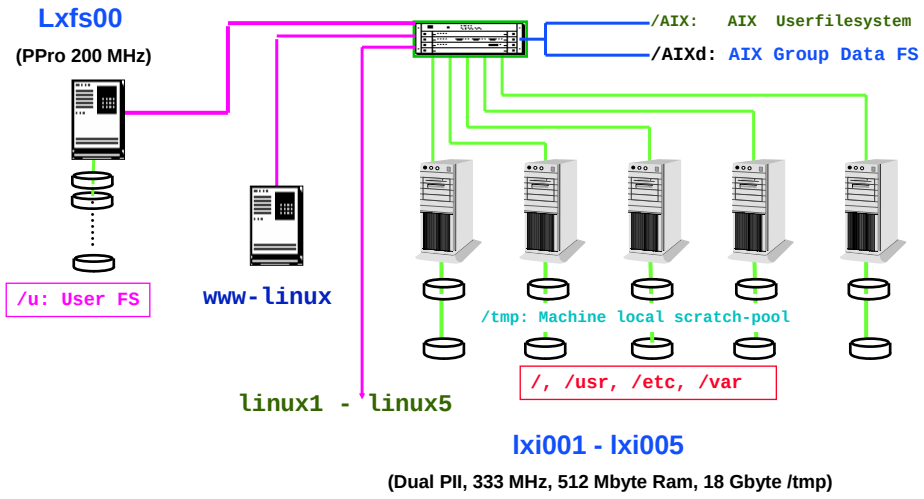
- kein Root Passwort an Users
- Ausnahme: Gruppenadministratoren auf ihrem Gruppenserver mit eindeutigen Regeln
- Support nur für GSI Farm Systeme
- User Desktops nur als Desktop Clients in die Farm integriert
- Zugriff auf HomeFS nur für Farm Systeme
- eingeschränkte Namesgebung der Farm Systeme
- FS (Data, ...) nur zentral, kene Crossmounts
- festgeschriebene PC Hardware
- kein Dualboot
- keine Notebooks für Linux

2. GSI Linux 1997/98

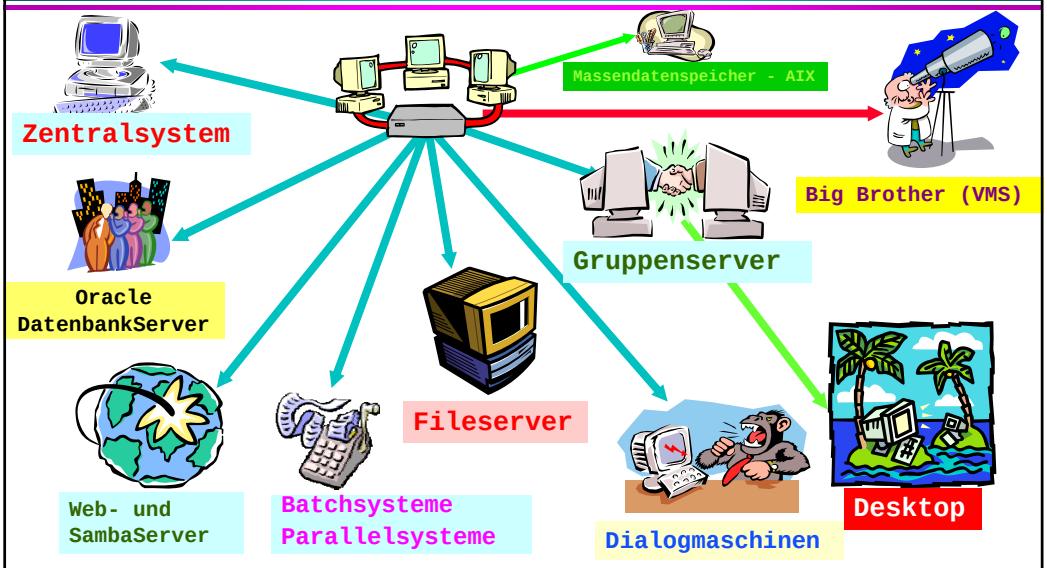
- Realisierung begann vor 3 Jahren
  - Auswahl - Hardware (5 + 1 PC)
    - Software
      - Distribution
      - Installierte Software (Compiler, ...)



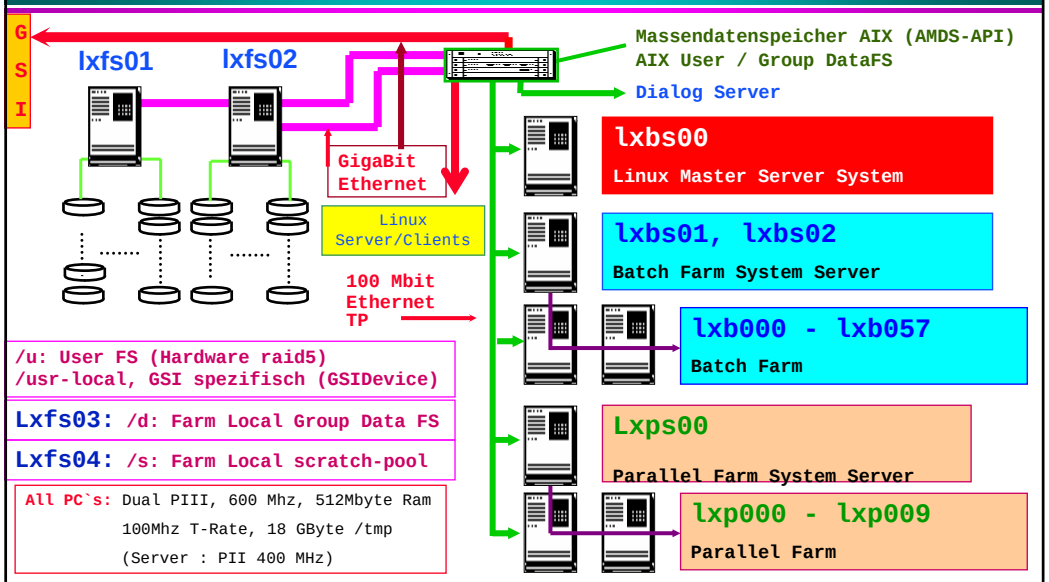
2. GSI Linux 1998/99



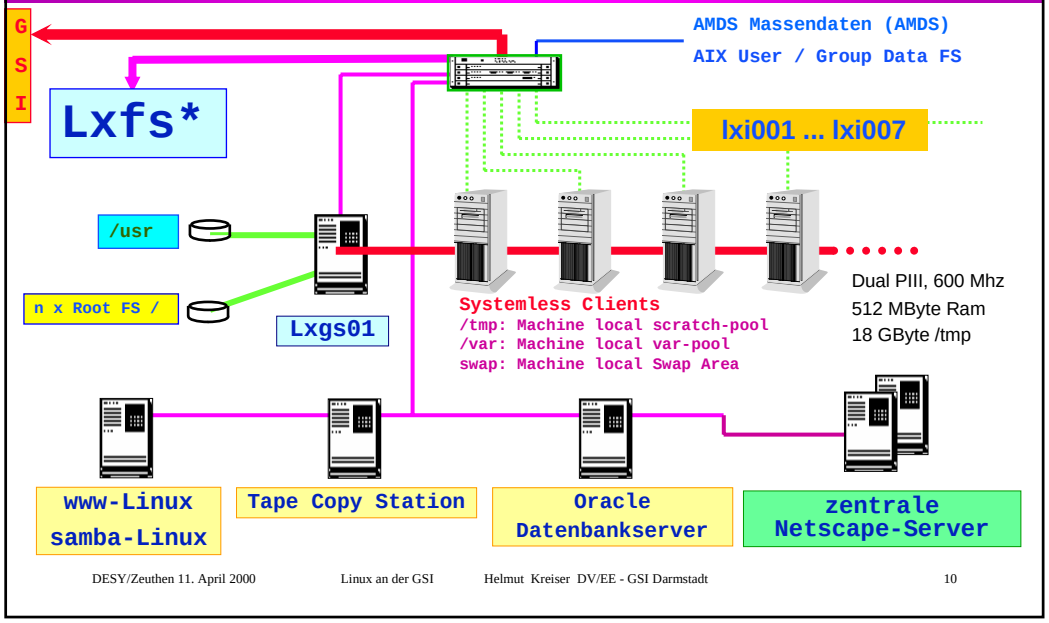
3. Die Linux Farm an der GSI



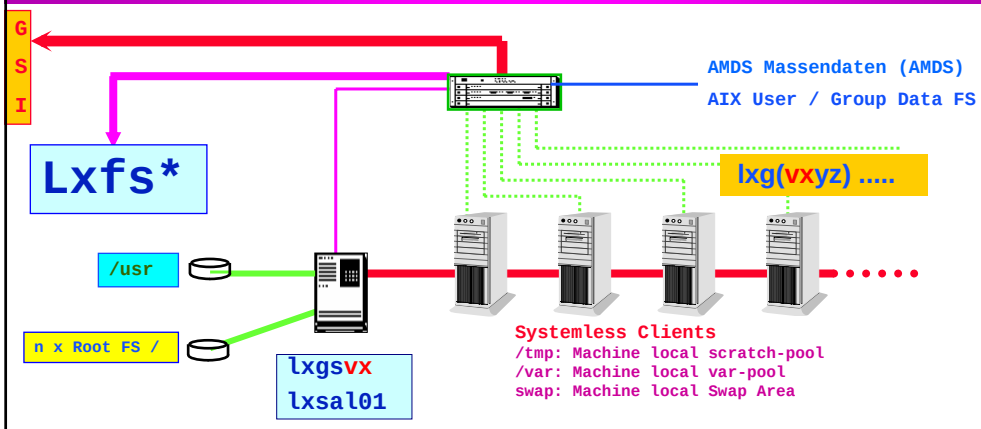
3. Zentrale Linux Server



3. Zentrale Linux Dialog Server



3. Linux Gruppen Server



lxgs00 = allgemeiner Server	lxgs04 = Hades
lxgs01 = RZ-Server	lxgs05 = Theorie
lxgs02 = FOPI	lxsal01 = Alice/NA49
lxgs03 = KP3	

3. Hardwaredefinition

Farm

- Asus Doppelprozessorboards
- Pentium PIII 600 Mhz
- 512 Mbyte RAM
- SCSI LvD (ausschließlich)
- ATI Graphik
- Intel Pro Netzwerkkarte (Twisted Pair)
- Gigabitkarten für Fileserver
- 18 GByte Platten

Desktop

- Asus Prozessorboards
- Pentium (PIII)
- xxx Mbyte RAM (256 Mbyte default)
- SCSI (ausschließlich)
- ATI Graphik
- Intel Pro Netzwerkkarte (Twisted Pair mit EthernetBox)
- 18 GByte

- keine Notebooks unter Linux

Farmumfang: 170 aktive Systems in der Linuxfarm
4 Fileserver + 1 Mastersystem + 11 Spezialsysteme
7 Gruppenserver
64 Batch- und Parallelsysteme
83 Desktopsystems

3. Linux Farm Hardware

Plattenbereiche an den Fileservern

- Storage:
  - einzelne Platten (spezielle System-Anwendungen)
  - Raid Array's (Compaq Storage Works Raid Array 3000)
    - Raid 5 (mit Spare Platte) für User FS
    - Raid 0 (Stripeing Set) für Daten-/Scratchpool
- /u - 4 Partitionen je 40 GByte als User-Homeverzeichnisse (160 GByte) (Erweiterung auf 7 x 40 = 280 Gbyte)
- /s - globale Scratch-Disk für alle (100 Gbyte) (Erweiterung: 200 Gbyte)
- /d - gruppenbezogene Datenpools (800 GByte) (Erweiterung: 1400 Gbyte)
  - RZ stellt Installation
  - Gruppen stellen Platten
  - kein zentrales Backup
    - /d/alice
    - /d/rz
    - ...
- /tmp system-lokaler Temp Bereich (max 18 GByte) (auf allen Rechnern)
- /loctmp lokaler Temp (max 15 Gbyte) Bereich (keine Clear-Policy)
- /www 1 Partition für User-Webpages (18 GByte) (Erweiterung>18 Gbyte)

3. Linux Farm Hardware

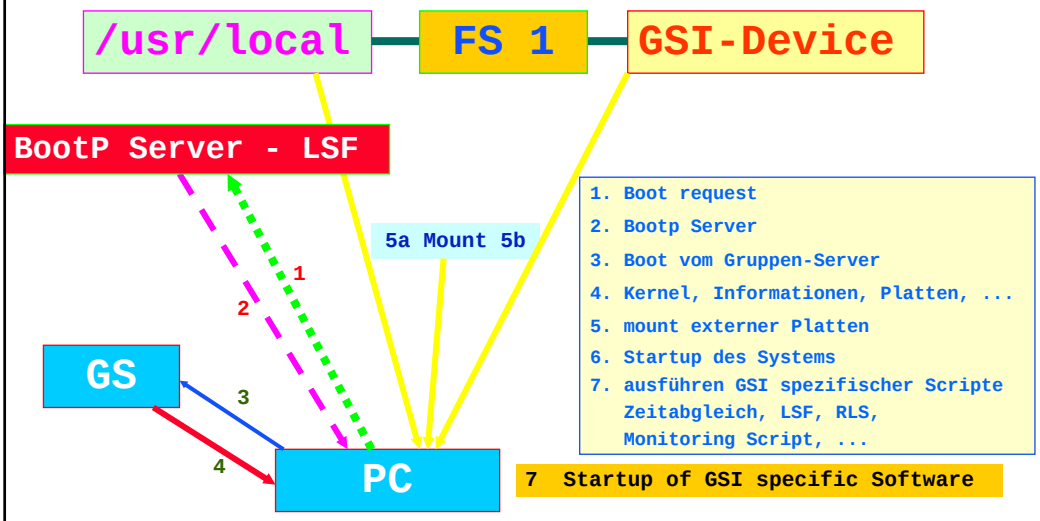
Plattenplatz: 3396 GByte (4216 GByte)

/u	: 160 GByte auf Hardware raid5 (max 280 GByte)
/usr-local	: 18 GByte
GSI/LSF Home	: 18 GByte
/d	: 800 GByte (max. 1400 GByte)
/s	: 100 Gbyte (max. 200 GByte)
/tmp	: 100 * 18 GByte (1800 GByte)
/misc	: 500 GByte (Backup-disks, Sytem-Disks, ...)

3. SystemServer/Systemless Clients

- Aufbau:
  - Systemserver
    - Host spezifische Root-FS (exportiert)
    - gemeinsames /usr
    - Installation über spezielle Scripte
  - Systemless Client
    - lokal:
      - /var
      - /tmp
      - swap
    - Installation manuell (Platte partitionieren, /var erzeugen)
    - Ziel: Bootfloppy mit Config-Scripten
  - Configurationsänderungen / Upgrades erfolgen nur auf den SystemServern
- Updatemöglichkeiten durch enge Verknüpfung der Farm (system-/gruppenabhängig)
  - einzelne Produktupgrades (automatische Prozeduren)
  - Farm Upgrade
  - Systemspezifisch: /Configuration/lppinst
- Vorteile
  - Updates erfolgen auf wenigen Maschinen (Server)
  - Fehlereingrenzung
  - zusätzliche Softwareinstallation auf Gruppenservern möglich

4. GSI - Device



4. GSI - Device

- Zentrale Platte
- verschiedene Verzeichnisse
  - 2 Konfigurationsfiles :
    - System-Gruppenzuordnung (Fileserver, XDialogserver, Xclient, ...)
    - Allgemeines Startupscript (Layered Products)
      - Zuordnung aller Hostnamen zu zusätzlich zu startenden Scriptnamen
      - laufzeiterzeugte, systemabhängige Startupscripts
      - LSF Startup, Zeitabgleich, Datenbankstartup, ...
  - Systemeigenes Startupscript
- zentrale Environmentvariablen (beim Login ausgeführt)
- zentral verwaltete Configurations-Files (/etc, Maps, ...)
- Maschinendirectories (Backup spezieller (Config-) Files)
- zentrale Steuerung für cron-Jobs (alle 30 min, täglich, wöchentlich, monatlich)
  - GSI spezielle Scripts
    - Config-Files Update (/etc, Maps, ...)
    - Softwareupdates
    - Prozessüberprüfung
- Monitoring (RLS, GSI Monitor Tool)

5. Software

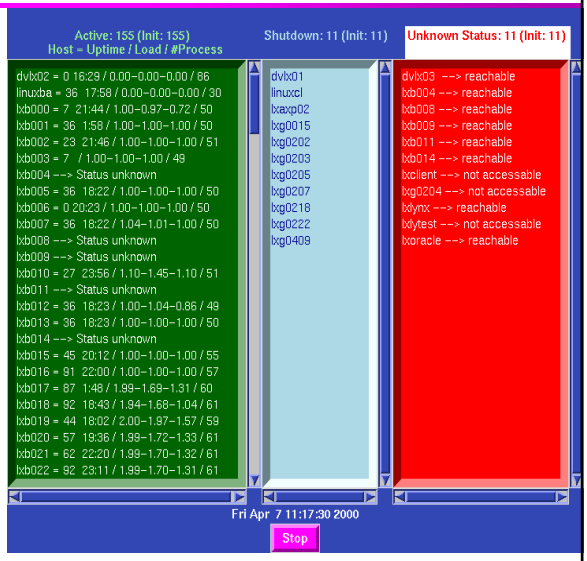
- Identischer Softwarestand innerhalb der Farm
  - Kernel 2.0.36/2.2.13, reduziert, geringe Modifikation
    # Prozesse, # pty's, nfs mount des root-FS
  - Standardcompiler (egcs: gcc, g++, g77, ..)
  - zugehörige Libraries, auch für Development
  - alle möglichen Tools (Editoren, Perl, Phython, ...)
  - Motif Run-Time auf allen Systemen
  - Motif Development auf zentralen lxi...
  - Pallas Compiler: Fortran, C, C++ (kommerziell)
  - IDL
- /usr/local/Linux:
  - Debian Distribution, erstellte Disk-Images (zum CD-Brennen)
  - KDE (incl. Source)
  - Linux Kernel Source, sowie angepaßte GSI-Kernel
  - sonstige Freeware
  - angepaßte GSI Prozeduren (ppp Zugang zur GSI über Modem)

6. Services

- Einteilung der GSI Linux Systeme:
  - Batch Systeme mit LSF (Batch Queueing System)
  - Desktop
    - zentrale interaktive Systeme
    - User-Desktopsysteme
- Interaktiven System (lxi001 - lxi007)
  - Keine Jobverarbeitung
  - nur submit nach LSF (Lizenzkosten)
- Batch Systeme (lxb000 - lxb056)
  - keine Login möglich, Status von Jobs kann über LSF abgefragt werden
- Web (www-linux)
  - Unterverzeichnisse im User HomeFS (web-docs, web-bin)
- Samba
  - Service Samba-Linux
- Mail
  - Mailzugriff über Netscape/Pine (IMAP-Protokoll) auf MS Exchange Service
- Backup
  - Full Backup monatlich auf Tape
  - Differential Backup nächtlich auf Platte (max. 7 Backups werden vorgehalten)
  - Vorbereitung ADSM Backup

6. Monitoring

- GSI own Tool
  - perl und perl/TK
  - Verzeichniseinträge der Maschinen GSI Device (alle 30 min active/shutdown/unknown und Date)
  - Abfrage von Systemparametern (uptime, load) über perl-socket
  - ping auf „unknown Status“ Systeme (reachable/not rea.)
- /var/log
  - lokal auf den Systemen
  - Backup alle 7 Tage (rotieren)
  - rls mit GUI (konfiguriert)
- SYSLOG Serielle Schnittstelle:
  - Console Manager (Alpha VMS)
- Video Switch System (kaskadiert)
  - für BIOS/Booten



7. Batch Queueing System (LSF)

- Version 3.2.2
- 57 PC’s für Batchverarbeitung (remote) (lxb000 - lxb056)
- 7 Maschinen zum Submitten/kontrollieren (lxi001 - lxi007)
- Auswahl der Hosts nach Load condition und resource requirements
- LSF benutzt Unix-Scripts
- max 2 Jobs pro Maschine
- max Anzahl von Jobs pro Benutzer ! (fair)
- Queues:
  - quick : max. 15 min, höchste Priorität, max. 3 Jobs
  - Jobs in anderen Queues werden angehalten
  - short : max. 1 h, hohe Priorität
  - research : max 5 Jobs, höhere Priorität, für spezielle Analysejobs
  - batch : kein Zeitlimit, normale Priorität
  - night : nur zw. 20.00 - 8.30, und am Wochenende
- Jobs mit teilweise Wochenlangen Rechnungen
- Job Outputs werden per Mail zugesandt, Output einschränken

8. Zahlen

```
Auslastung Batchfarm      : im Mittel 60 - 70 %
                          : zeitweise 100 %
Datentransfer Taperobot <-> Linuxfarm Rechner (Socket API)
                            mittlere Transferrate : ~10 Mbyte/sec
Gesamttransfer 1999  Robot -> Tape-Rechner : 52,5 TByte
                                  Linux : 76 %
                                  AIX   : 14 %
                                  VMS   : 10 %
                     Rechner -> Tape-Robot : 4 TByte
                                  Linux : 59 %
                                  AIX   :  9 %
                                  VMS   : 32 %
Auslastung Home FS        : ca. 38 % - 99 %
Auslastung Scratch FS     : 100 %
Auslastung Daten FS       : ständig zunehmend (3% - 91%)
# Prozesse auf zentralen interaktiven Hosts : > 256 (max 4096)
# Users                   : > 20 - 30 pro zentraler interaktiver Maschine
Max Ausbau der Farm:       7 Gruppenserver (25 Clients) : 175
                           2 Batchserver   (35 Clients) :  70
Ausbau ohne weitere Änderung: 170 -> 265 Systeme
```

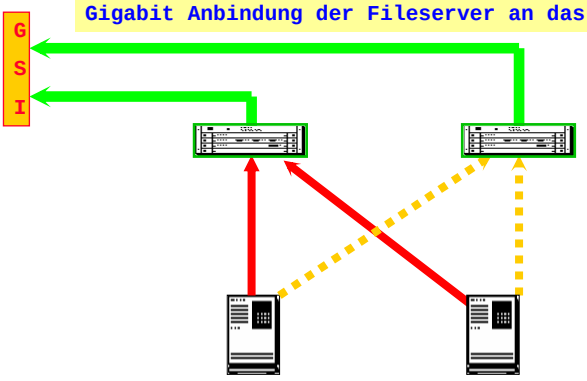

9. Probleme auf Linux

- NFS Server:
  - von Lynx wurden keine Files mehr geschrieben, max 4KByte Blocks !
  - Patch nach über 5 Monaten ! (NFS-Server 2.2beta37, immer noch nicht benutzbar)
  - --> alter NFS-Server
- Probleme im SMP (Interrupt) Handling bei Dual-CPU‘s (2.0.36)
  - starke I/O Last bei Plattenzugriffen (read/write gleichzeitig)
  - u.U. hängende Prozesse
- Zeitdifferenzen (trotz ntp-Abgleich; Network Time Protokoll)
- Filegröße max. 2 GByte
- NFS Hänger (u.U. mit Abstürzen) bei Zugriff auf die Server (hohe Netzwerk-/Serverlast)
- Abstürze ohne erkennbare Gründe (Memory Probleme ?)
- lange Filesystemchecks bei Abstürzen der Fileserver (GSI: bis zu 3 Stunden)
  - Journal File System
- Fehlende Farm-Überwachnungstools
  - GSI Own Monitoring Tool
  - Scout ?

10. Linux Farm Future

- zusätzliche Services
  - Printservice (von AIX)
  - NIS+ (von AIX)
  - Nameservice ??

Gigabit Anbindung der Fileserver an das GSI Backbone



Fileserver: Dual Port Gigabit Karten mit automatischen Failover

10. Linux Farm Future

/u, /d Group-Data, /s

Zusätzliche Plattenkapazität

Entwickeln/benutzen von Farmüberwachungstools evtl. mit Konfigurierungsmöglichkeiten

All PC`s: Dual Px, xxx Mhz, >=512MbyteRam
100Mhz T-Rate, >=18 GByte /tmp

lxps00, ...?
Parallel Farm System Server

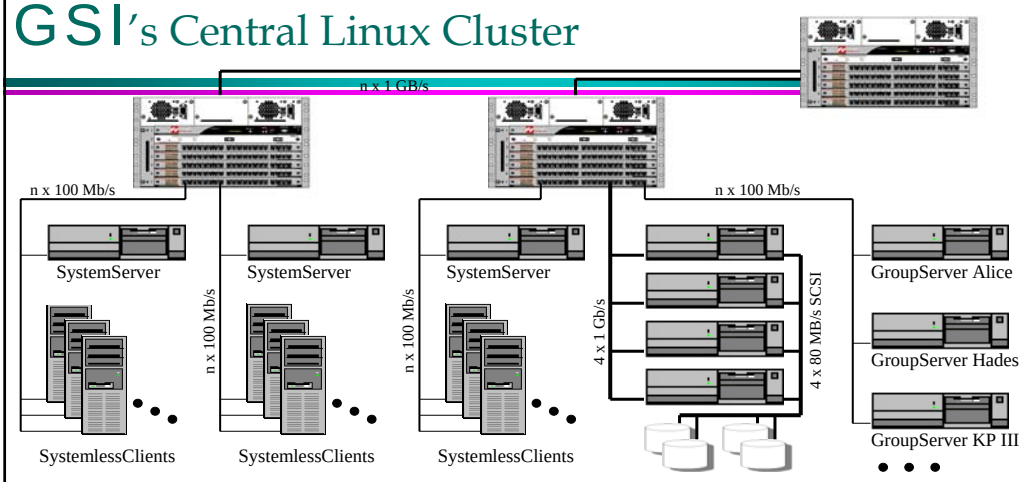
lxp000 - lxpxyz
Parallel Farm: Myrinet Netzwerk, möglicher Ausbau auf Alpha Linux

lxi001 - lxixyz
Dialog Farm

lxgs00 - lxgsxy
Group-Server

lxg0000 ...
Group Desktop PC‘s

GSI's Central Linux Cluster



	Interactive	Batch	FileServer	GroupSystemServer
Server	1 Dual Pentium II (2 CPUs), 400 Mhz 512 MB, 4.3 GB	2 Dual Pentium II, (4 CPUs) 400 Mhz, 512 MB, 4.3 GB disk	4 Dual Pentium II, (8CPUs), 333 Mhz; 512 MB, 4.3 GB	7 Dual Pentium II, (14 CPUs), 400 Mhz, 512 MB, 4.3 GB
Client	7 Dual Pentium III (14 CPUs), 600 Mhz 512 MB, 18 GB	70 Dual Pentium, (140 CPUs) 600 Mhz, 512 MB, 18 GB disk	HomeArea: 400 GB DataArea: 1400 GB ScratchArea: 200 GB	Supports up to 250 systemless desktop clients at GSI

H.Kreiser, A.Mann, M.Richter, W.Schiebel 01.02.2000