UNICOS/mk

skalierbares
Microkernel-Betriebssystem

für

CRAY Research
massiv-parallele Rechnersysteme

OS: Attached Processor Modell

- **Vorteile**

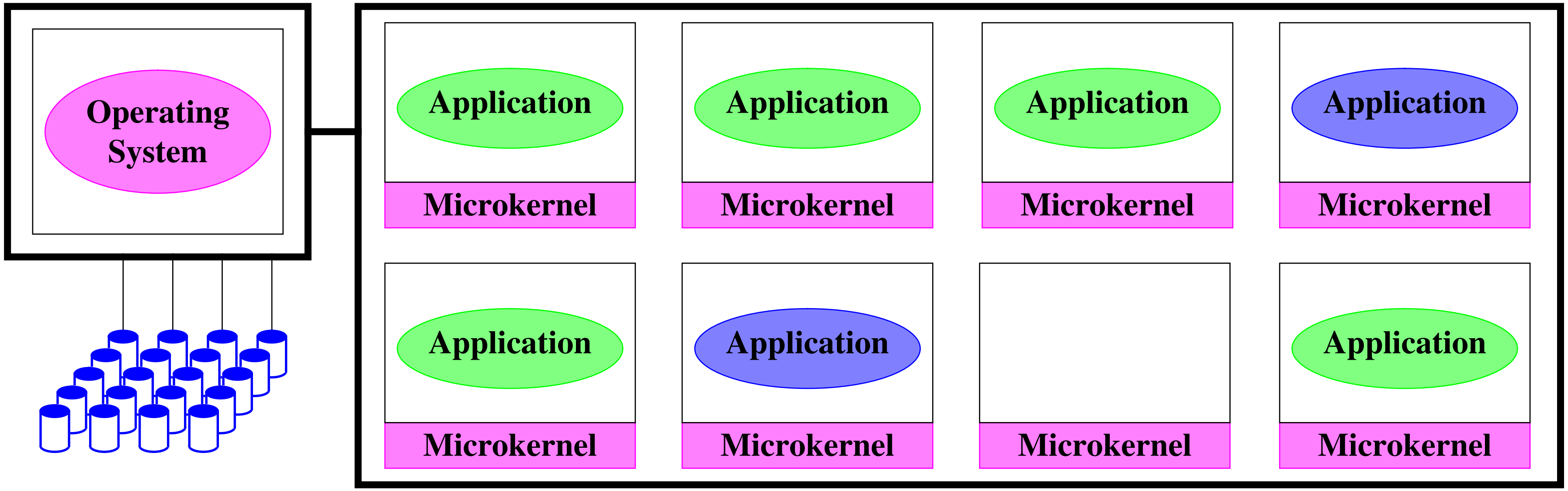
  Direkter Zugriff auf alle Ressourcen (Single System Image)

  Ein Prozeßraum

- **Nachteile**

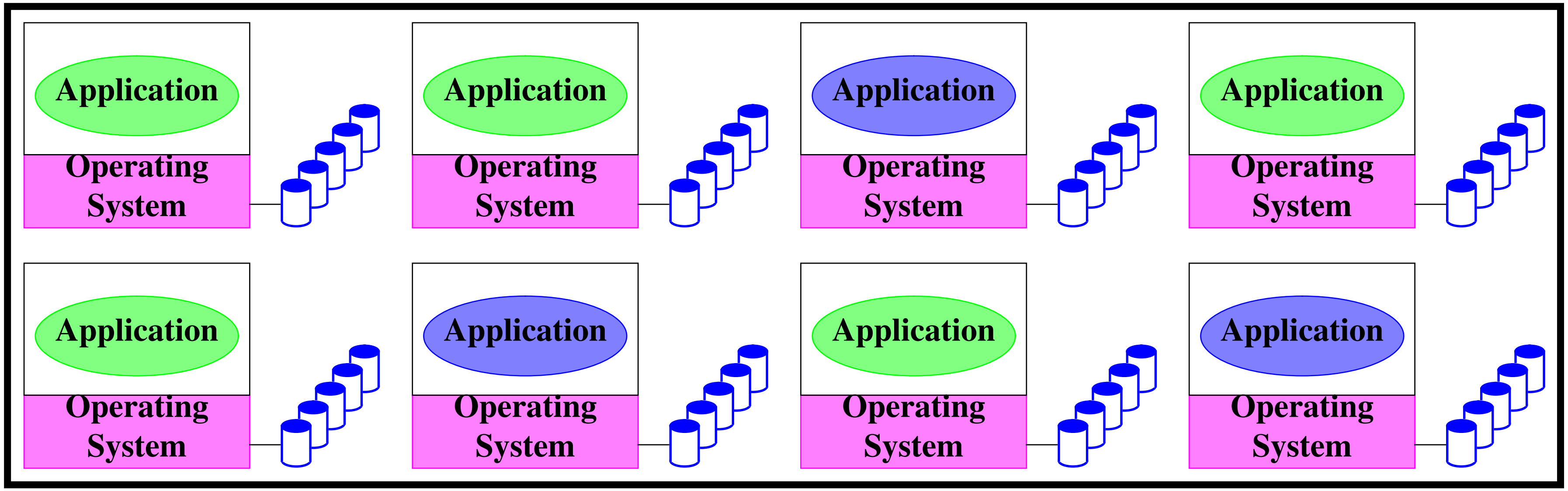
  Begrenzte Skalierbarkeit, speziell I/O

  Starker OS-Server benötigt



OS: Workstation Farm Modell

- **Vorteile**
  - Einzelprozesse sehen nur geringe Latenzzeit zum lokalen Betriebssystem
  - Schneller lokaler I/O
- **Nachteile**
  - Prozeß-Management und System Administration sehr aufwendig
  - Hoher Overhead durch Replizierung



OS: Verteiltes Betriebssystem Modell

- **Vorteile**

  Emuliert ein 'Attached Processor Model' (Single System Image)

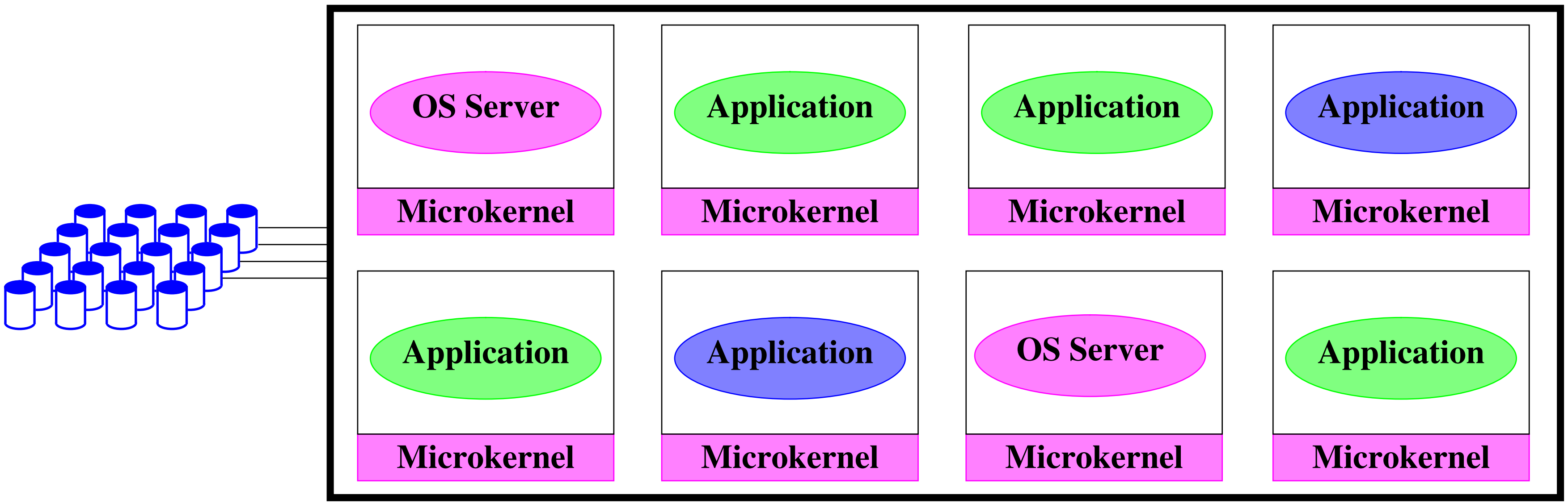
  Geringer Speicherplatzbedarf

  Skalierbarkeit

- **Nachteile**

  Einzel-PE Anwendungen nicht so schnell wie im Workstation Farm Modell

  Geringe Latenzzeit zwischen PEs nötig



UNICOS/mk - Überblick

- **Für die große Anzahl von Prozessoren einer MPP-Architektur ist es erforderlich, die Funktionen des Betriebssystems zu verteilen**
- **Das neue, auf dem CHORUS Microkernel basierende Betriebssystem wird zur Unterscheidung mit UNICOS/mk bezeichnet**
- **Gründe für CHORUS (anstelle von MACH, Amoeba, etc.)**

  Bestes Speicher Modell (Unterstützung des Real-Speicher-Konzepts)

  Existierende Methoden zur Performance Verbesserung

  Light weight rpc (user space) - full weight rpc (kernel space)

- **Entwicklungsziele für das umstrukturierte System:**

  Erhaltung der UNICOS Benutzerschnittstelle (z.B. Login, Prozeßverhalten)

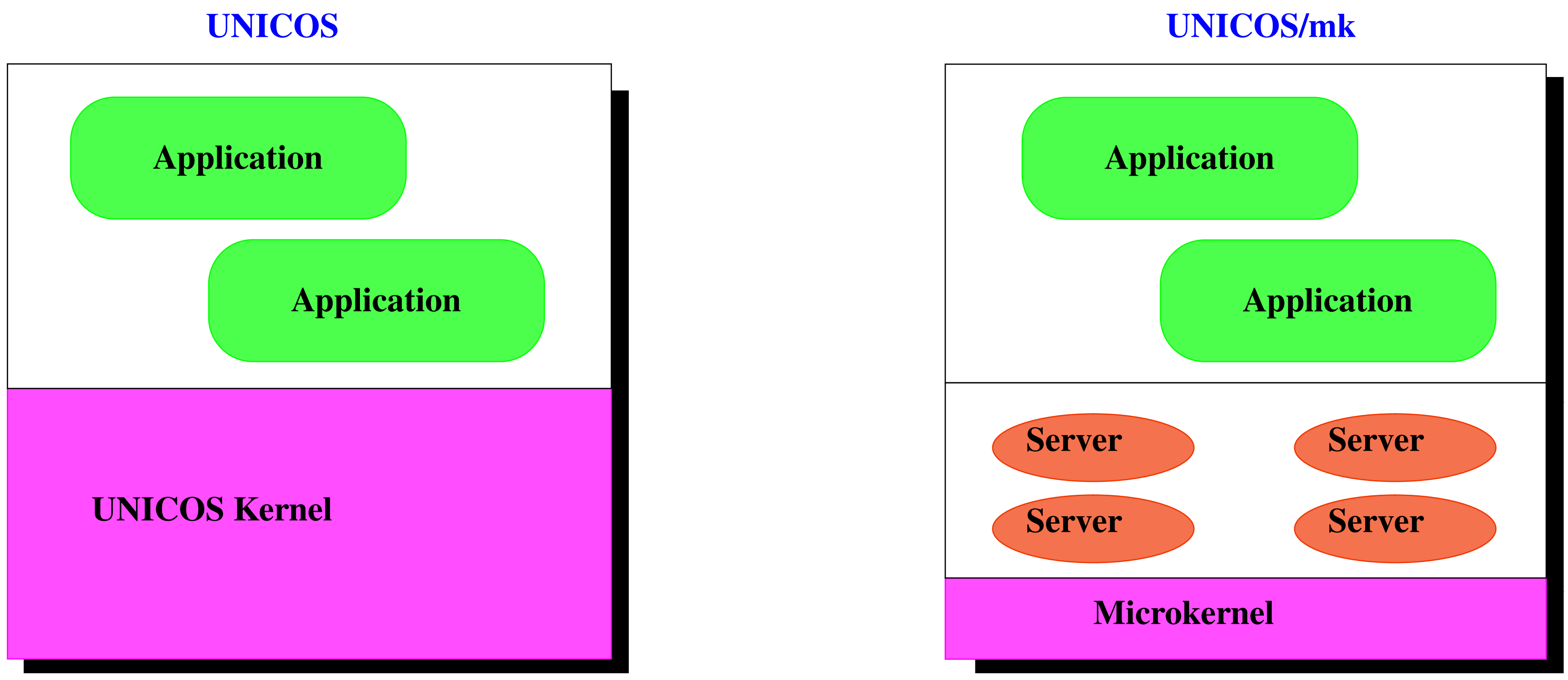
  Einheitliche Schnittstelle aus Sicht einer Anwendung (System Call Ebene)

  Einheitliche Sicht für den Systemadministrator (z.B. Ressourcen-Vergabe)

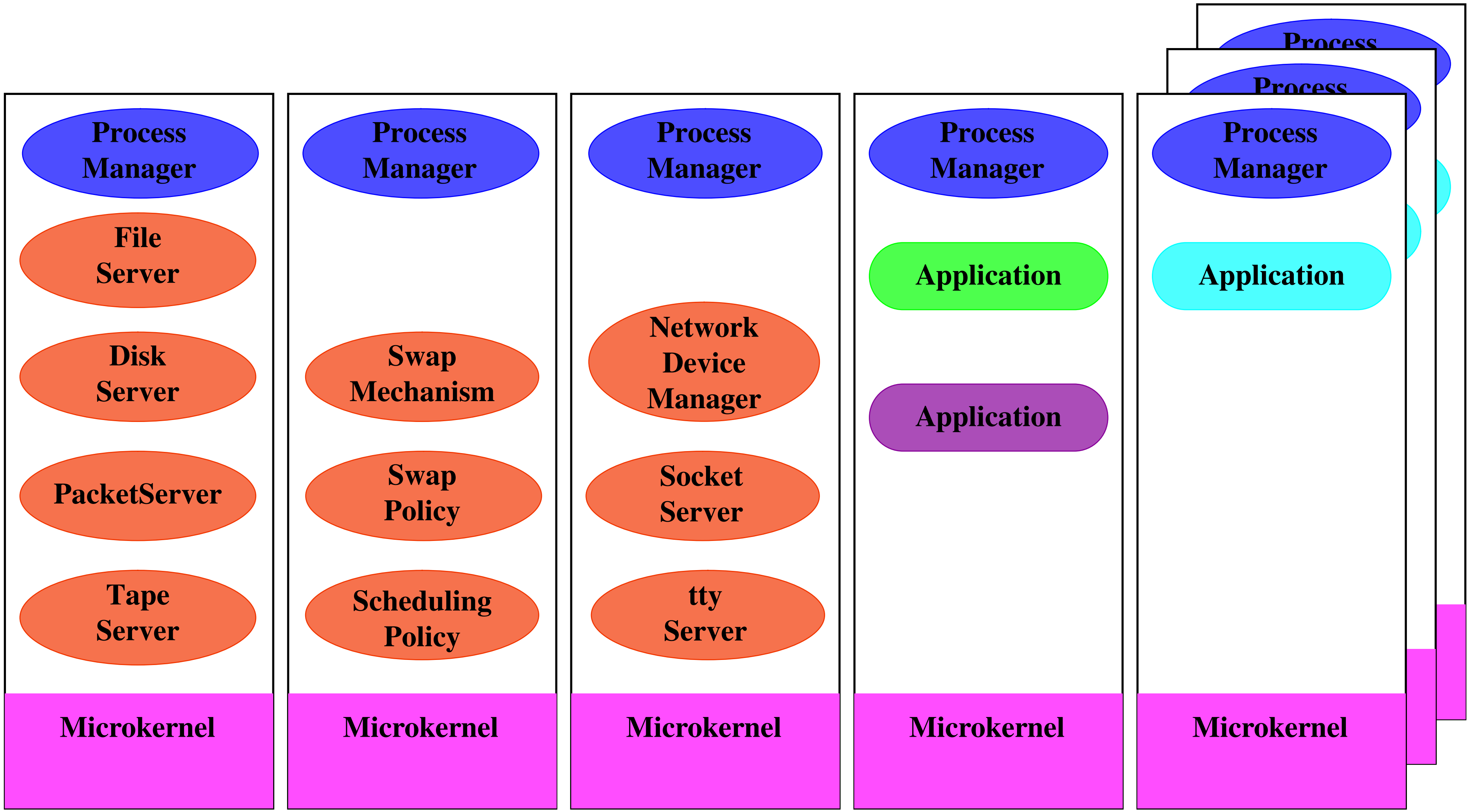
  Wiederverwendung des vorhandenen UNICOS Kernel Codes

UNICOS/mk - Überblick

- **Ein kompakter Microkernel läuft auf jedem Knoten und implementiert nur die wichtigsten Basisfunktionen wie Speicherverwaltung, Interprozeß-kommunikation, Basis-I/O und die Hardware-abhängigen Funktionen**
- **Alle anderen Betriebssystemfunktionen sind modular auf sogenannte Server verteilt**



Einige Komponenten des Betriebssytems UNICOS/mk

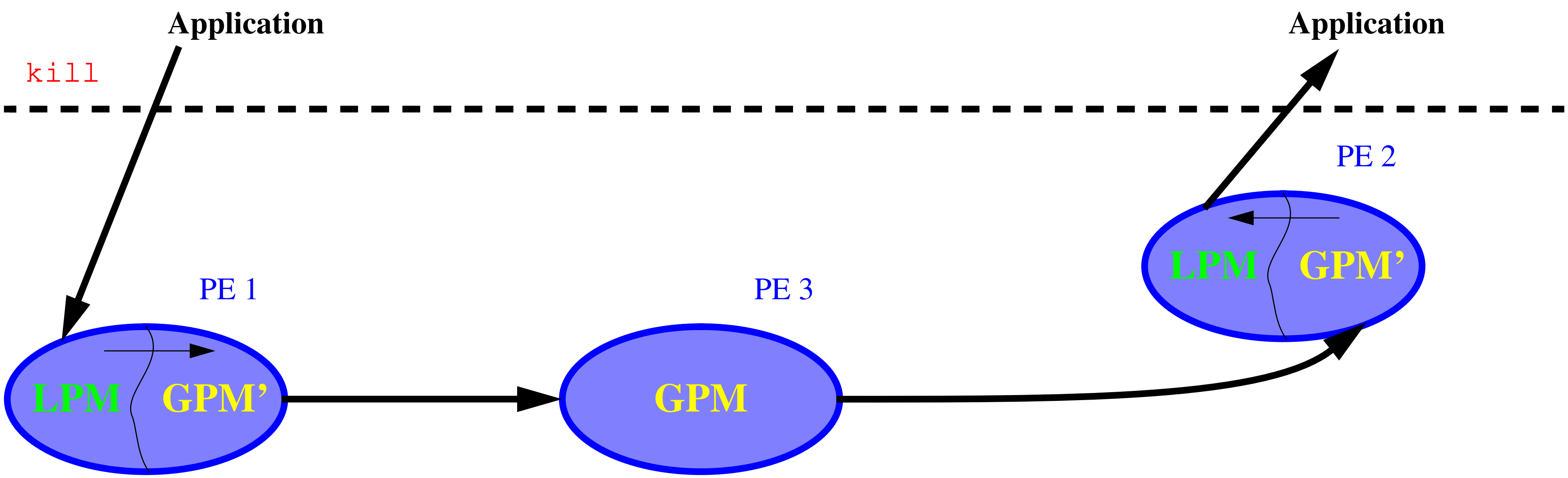


Verteiltes Betriebssystem UNICOS/mk

- **Process Manager auf jedem PE bildet die Schnittstelle zwischen Anwendung und Microkernel**

  Ein globaler Process Manager (GPM) verfolgt und koordiniert alle globalen Aktivitäten; liegt auf einem der OS Server PEs

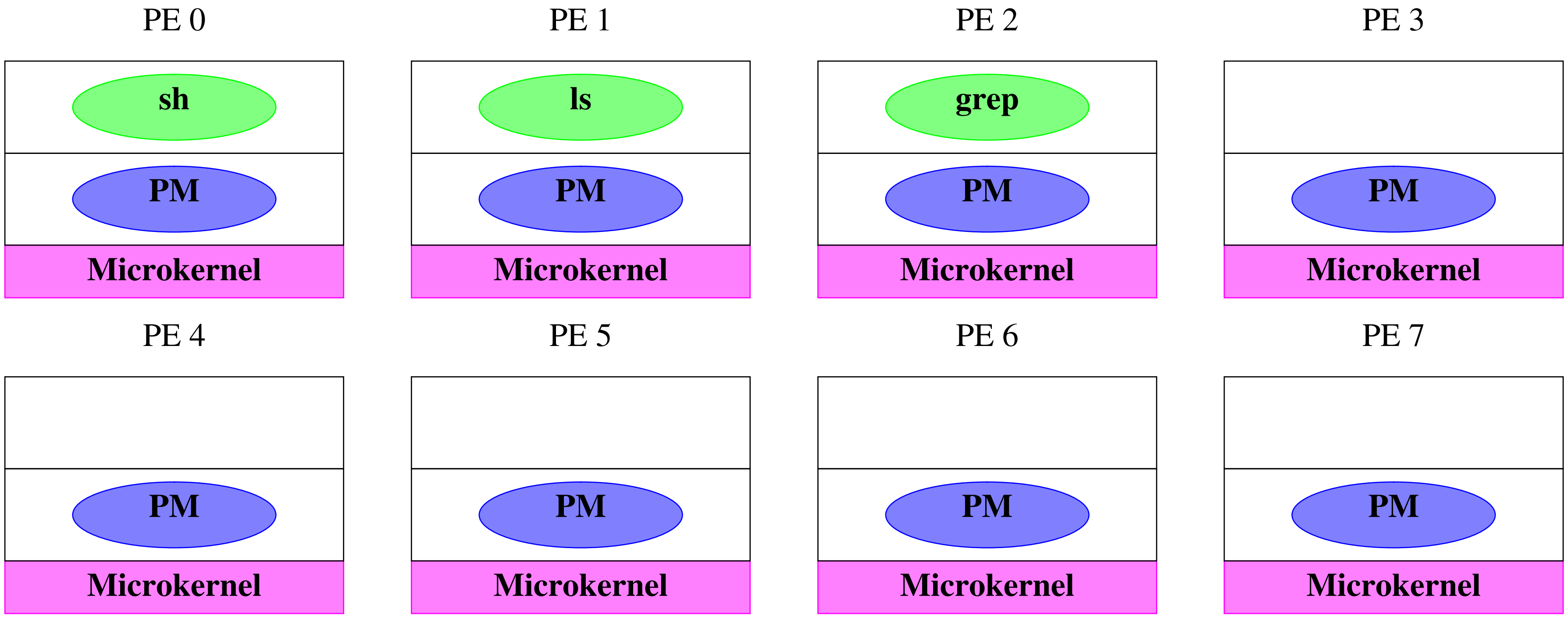
  Lokaler Process Manager auf jedem PE mit einem lokalen (LPM) und globalen Teil (GPM’)



UNICOS/mk Single System Image

- **Beim Einwählen in das System (Login-Prozeß), wird eine Shell gestartet**
- **Die Shell sowie alle weiteren Prozesse können dabei auf unterschiedlichen PEs des MPP-Systems ablaufen (least memory fit)**
- **Eine parallele Anwendung stellt lediglich eine Sonderform dar**

Komando: **ls | grep x**



UNICOS/mk Single System Image - PE-Klassen

- **Kommando PEs**

  Benutzer-Kommandos (z.B. Login Shell, Compiler)

- **Applikation PEs**

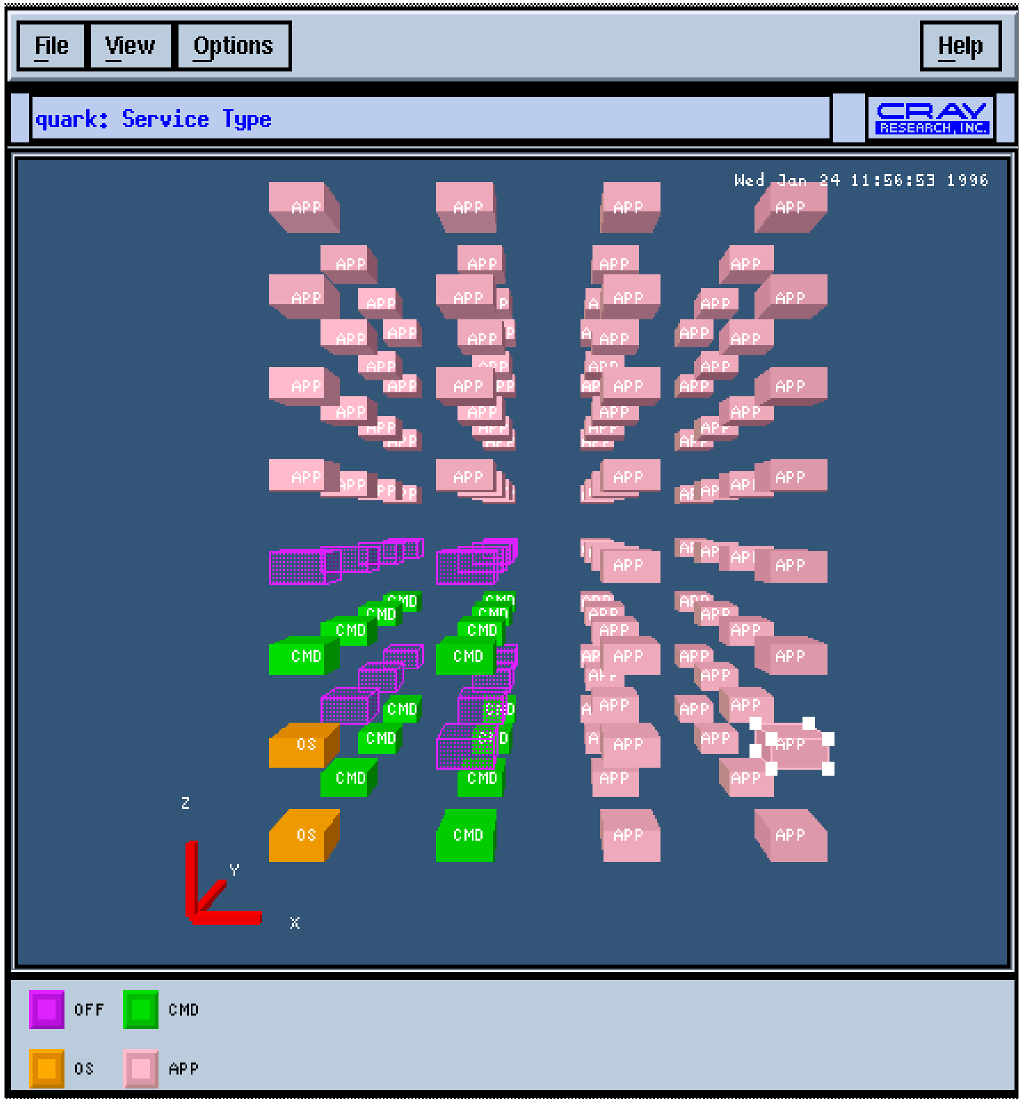
  Ausführung der parallelisierten Anwendung

- **OS PEs**

  Zur Ausführung der Server-Prozesse ohne zusätzliche Benutzer-Kommandos

- **Redundante PEs**

  Gleichsam als ’Hot-Spares’ im System vorhanden, um die Aufgaben ausfallender PEs zu übernehmen



UNICOS/mk Single System Image - Scheduling

- **Ressourcen-Attribut, die Charakteristik eines jeden PEs**

  Beinhaltet Faktoren wie minimale oder maximale PE Anzahl,Speichergröße, Taktfrequenz, Bediener-Typ (z.B. batch, interaktiv, rsh), UID und GID

  Die Spezifikation eines Ressourcen-Attributs limitiert die Prozesse, die auf dementsprechenden PE ausgeführt werden können

  Dieses Attribut wird vom Systemadministrator vergeben

- **Applikations-Attribut, Anforderungen oder Präferenzen der Anwendung**

  Kann vom Benutzer explizit oder implizit über den Benutzer-Eintrag in der UDB (User Data Base), bzw. über die NQS-Klasse gesetzt werden

  Beinhaltet Faktoren wie Speicherplatzbedarf, Anzahl PEs und Bediener-Typ

- **Präzedenz-Attribut, das vom globalen Resource-Manager anhand der beidenanderen berechnet wird**

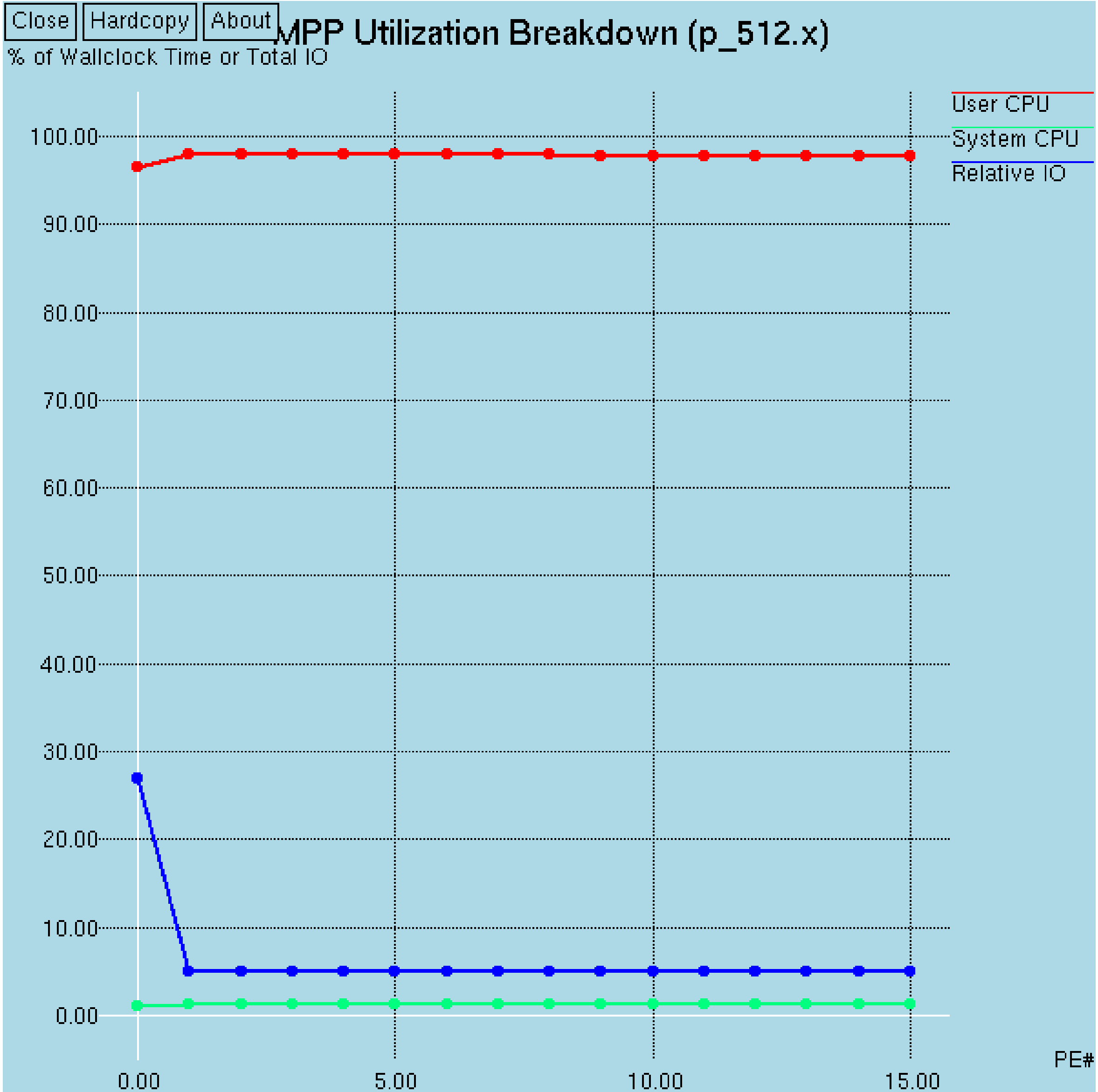
  Beinhaltet Faktoren wie Speicherplatz-Zuteilung (best fit), Zusammenhängigkeit sowie optimale PE-Zuteilung innerhalb einer Region

UNICOS/mk Accounting: ja -chl

```
Command  Started  Elapsed User CPU  Sys CPU Rmt CPU I/O Wait CPU MEM  I/O WMem Mwords  Log I/O Memory
  Name      At    Seconds  Seconds  Seconds Seconds  Seconds Avg Mwds Avg Mwds Xferred Request HiWater
======== ======== ======= ========= ======= ======= ======== ======== ======== ======= ======= =======
# CFG    09:56:26  System:  sn6546 hubble 1.0.0 chorus CRAY T3E
ja       09:56:25    0.97      0.05    0.13   0.02      0.32     0.29     0.28   0.002       1     592
grep     09:56:26    0.25      0.00    0.07   0.01      0.10     0.26     0.27   0.000       3     263
target   09:56:27    0.25      0.00    0.04   0.01      0.04     0.25     0.28   0.000       2     369
p_512.x  09:56:27   61.28    959.22   11.92   0.38      6.60     1.29     0.86   2.563      70   43216
# MPP
# MPP Usage     PE's= 16, BESU's= 1, PE Memory High Water= 1.32 MWords
# MPP
# MPP Statistic Minimum Maximum Average Std. Dev.
# MPP           ======= ======= ======= =========
# MPP             59.17   60.04   59.95      0.21   Seconds User CPU Time
# MPP              0.65    0.80    0.77      0.04   Seconds System CPU Time
# MPP              0.12    0.69    0.16      0.14   MWords transferred
# MPP
# MPP Clusters    59.17   59.17   59.17      0.00   Seconds User CPU Time (No. of PEs in Cluster : 1)
# MPP             59.95   59.97   59.97      0.01   Seconds User CPU Time (No. of PEs in Cluster : 7)
# MPP             60.03   60.04   60.04      0.01   Seconds User CPU Time (No. of PEs in Cluster : 8)
# MPP
# MPP Clusters      NA      NA      NA        NA    Seconds System CPU Time
                                                    (*** No Cluster could be identified ***)
# MPP
# MPP Clusters     0.12    0.13    0.12      0.00   MWords transferred (No. of PEs in Cluster : 15)
# MPP              0.69    0.69    0.69      0.00   MWords transferred (No. of PEs in Cluster : 1)
```

UNICOS/mk Accounting

- **Eingabe für xgraph zur anschaulichen Darstellung der Verteilung von User CPU Zeit, I/O-Wait Zeit, System CPU Zeit, über alle PEs paralleler Anwendungen**
- **ja -x > file**
- **xgraph file**

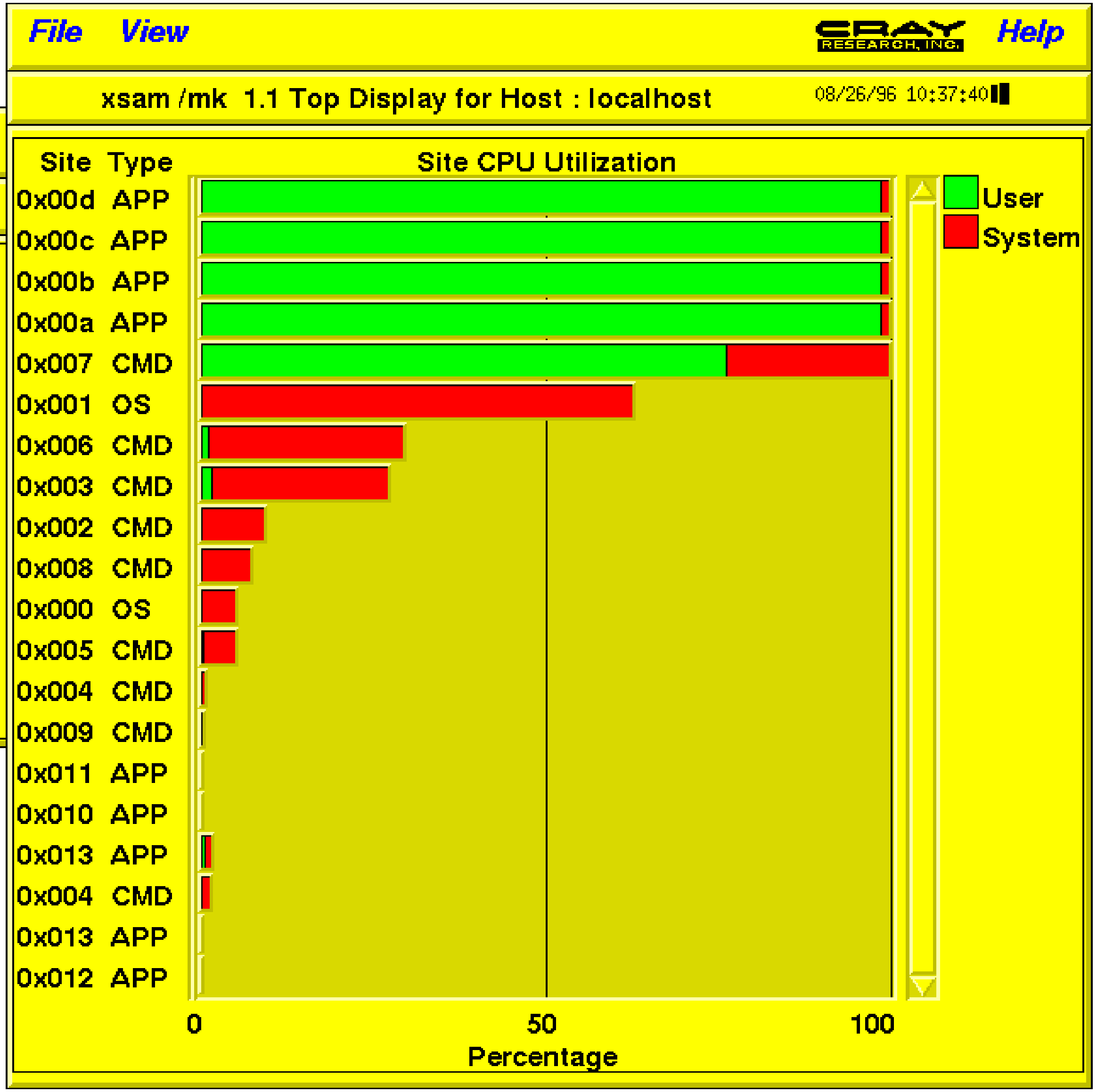


UNICOS/mk System Activity Monitor - xsam

File

xsam /mk 1.1 Configuration Display for Host : localhost

Machine	CRAY T3E	Mainframe Type	CRAY-T3E
Node	hubble	System	sn6546
Release	1.0.0	Version	rkb
No. of Sites	20	No. of OS-Nodes	2
No. of CMD-Nodes	8	No. of APP-Nodes	10
No. of Disks	32	No. of Mounts	0
No. of Tapes	0		
Physical Memory	2560 MBytes	Swap Memory	0 MBytes



A Silicon Graphics Company

UNICOS/mk System Konfiguration: grmview -l

- PE - Typ mit Attribut für minimale und maximale Anzahl
- Attribut für Anzahl Applikationen pro PE
- Verfügbarkeit (+) oder nicht (-)
- Logische und physikalische (x, y, z) PE Nummer
- Verfügbarer (abzügl. System) und freier Speicherplatz für Anwender
- Prozessorgeschwindigkeit in MHz

```
PE Map: 8 PEs configured
            Ap. Size  Number Aps.          <<<< Lists >>>>>
 Type    PE  min   max running limit    Label svc uid gid acid  x  y  z Clock UsrMem FreMem
+  OS     0    0    0       0     0       -     -   -   -    -  0  0  0   300     69     55
+ CMD     1    1    1       0 unlim       -     -   -   -    -  1  0  0   300    117     17
+ APP     2    2    6       0     1       -     -   -   -    -  0  1  0   300    120    120
+ APP     3    2    6       0     1       -     -   -   -    -  1  1  0   300    120    120
+ APP     4    2    6       0     1       -     -   -   -    -  0  0  1   300    120    120
+ APP     5    2    6       0     1       -     -   -   -    -  1  0  1   300    120    120
+ APP     6    2    6       0     1       -     -   -   -    -  0  1  1   300    120    120
+ APP     7    2    6       0     1       -     -   -   -    -  1  1  1   300    120    120
Exec Queue: 1 entries total. 1 running, 0 queued
  uid   gid  acid    Label Size BasePE
 1308 11834     0     -       4      2
```