

Speicherverwaltung

Betriebssysteme I WS 2006/2006



Jörg Kaiser
IVS – EOS

Otto-von-Guericke-Universität Magdeburg

Speicherverwaltung

**Wozu braucht man eine Speicherverwaltung?
Welche Eigenschaften sollte ein Speicher haben?**

- **unendlich groß,**
- **unendlich schnell,**
- **unendlich billig,**
- **nichtflüchtig,**
- **kein Effekt durch konkurrierende Zugriffe,**
- **Schutz vor fehlerhaften Zugriffen.**

Die Speicherverwaltung hat die Approximierung dieser Eigenschaften zum Ziel !



Speicherverwaltung

Die Kosten-Leistungs-Perspektive

sehr teuer

sehr billig



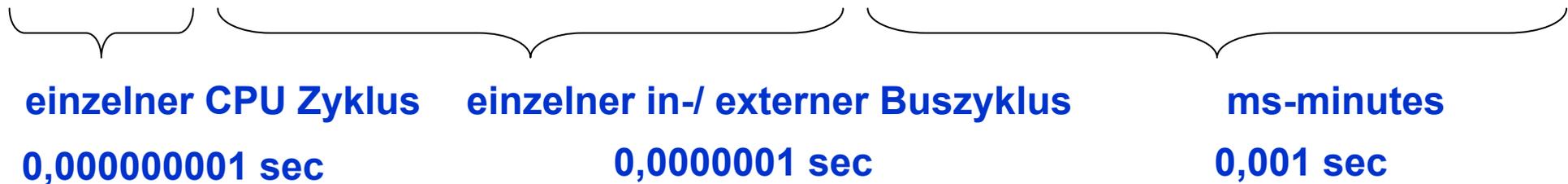
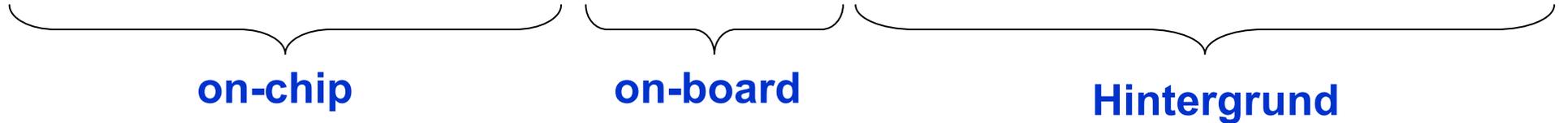
sehr schnell

schnell

langsam

sehr langsam

registers scratchpad caches RAM ROM Disk DVD-ROM CD-ROM Tape



Speicherverwaltung

Register
Scratchpad

wird explizit vom Programmierer/Compiler kontrolliert

Cache

Hardware-Kontrolle

RAM

RAM
Abstraktion

ROM

flüchtig

Disk

DVD-ROM

Dateiabstraktion

CD-ROM

persistent

Tape

unter Kontrolle des Betriebssystems



Speicherverwaltung

Themen für die Speicherverwaltung:

★ **Relokation**

★ **gemeinsame Nutzung**

★ **Zugriffsschutz**

resultiert aus dem Zugriff durch mehrere Prozesse/Programme

★ **Transparenter Zugriff über eine Hierarchie von physischen Speichern hinweg**

logische Organisation

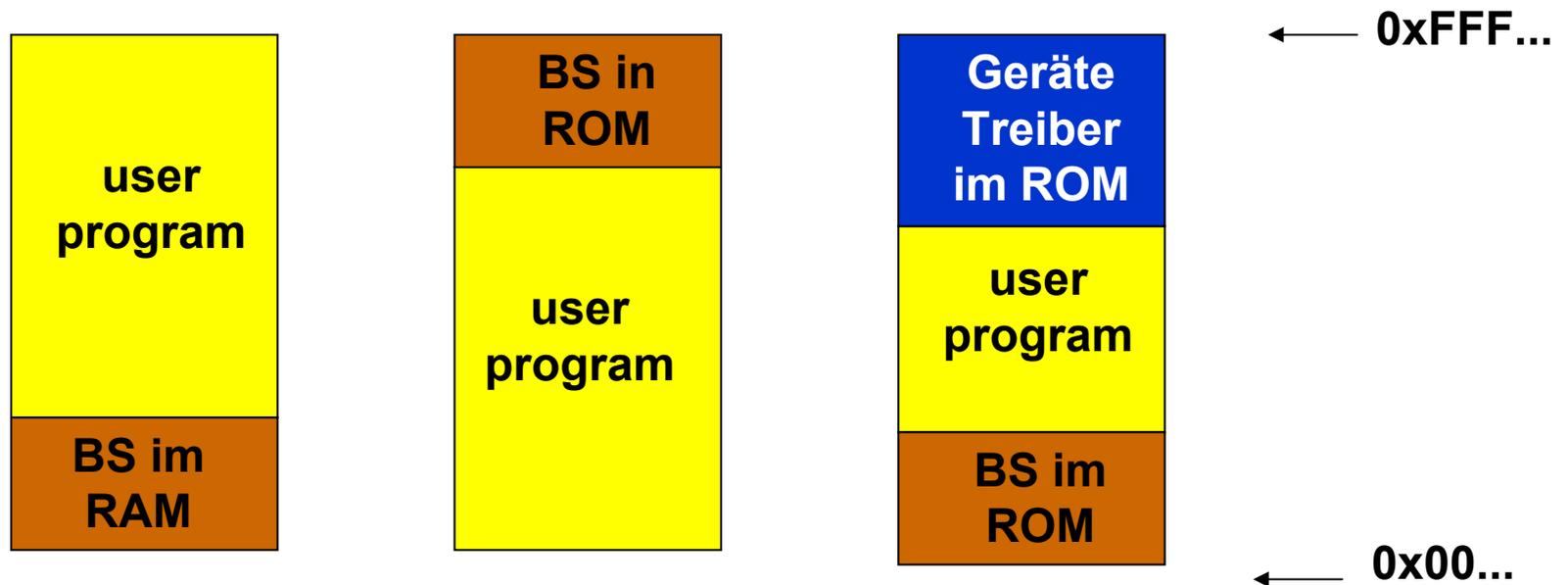
physische Organisation



Multiprogramming mit Speicherpartitionen

Statische Partitionierung:

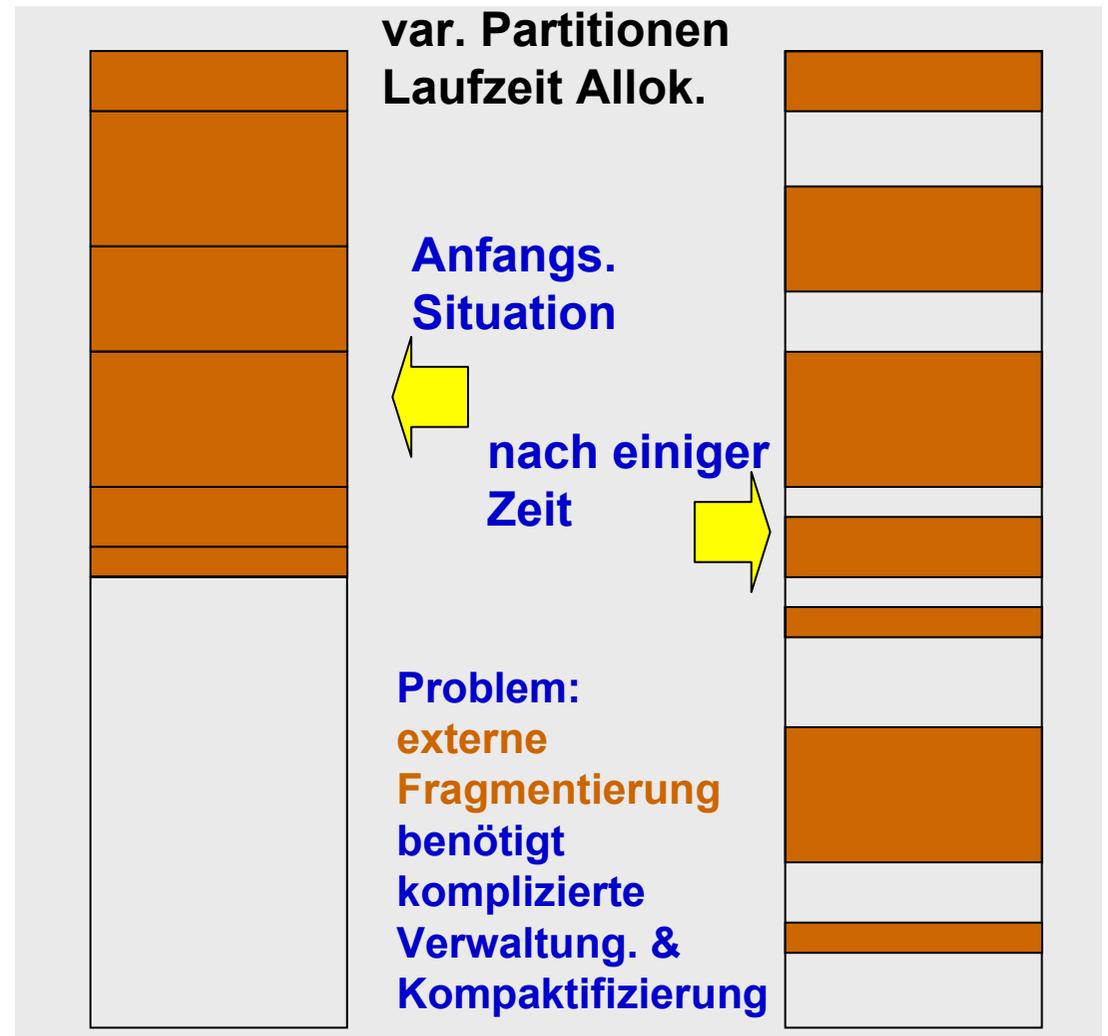
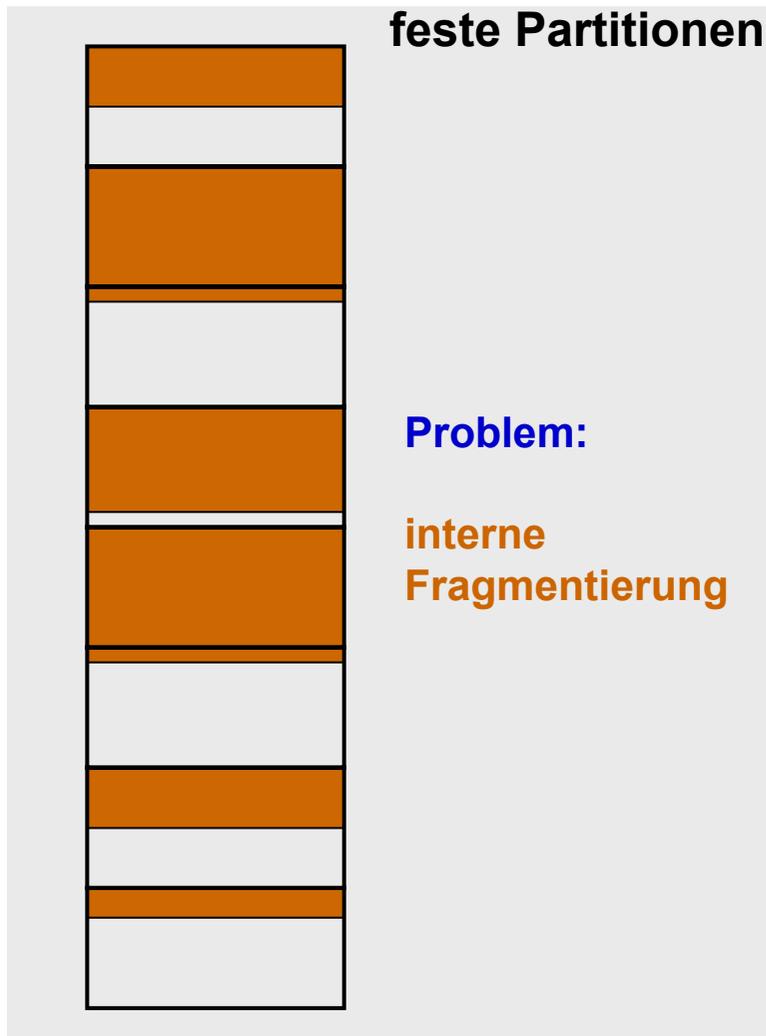
wird heute in einigen eingebetteten Systeme benutzt, z.B. Palmtops, PDAs, etc.



Eigensch.:	Größe der Speicherblocks	Zeitpunkt der Allokation
	feste Größe variable Größe	off-line (statisch) zur Laufzeit (dynamisch)

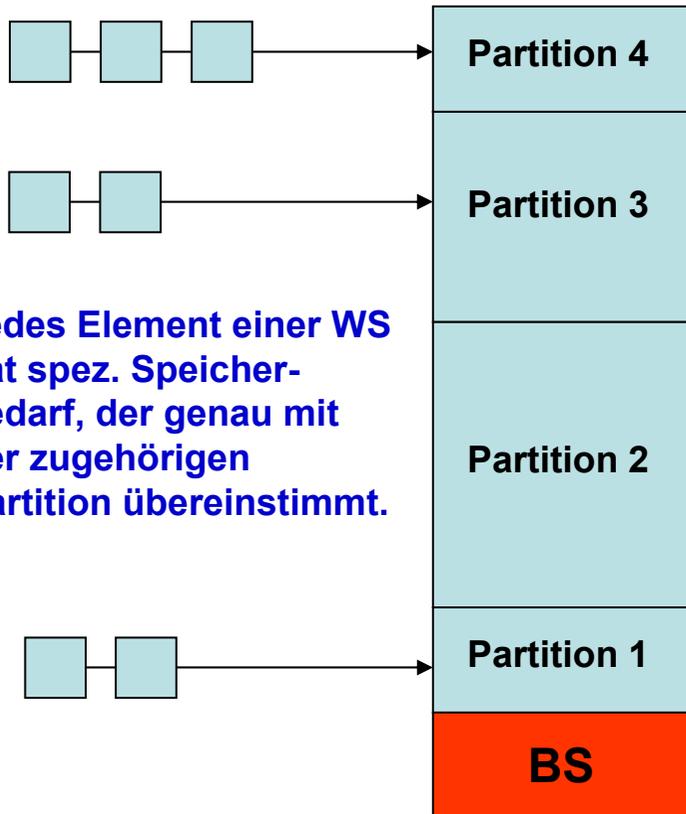


Multiprogramming mit Speicherpartitionen



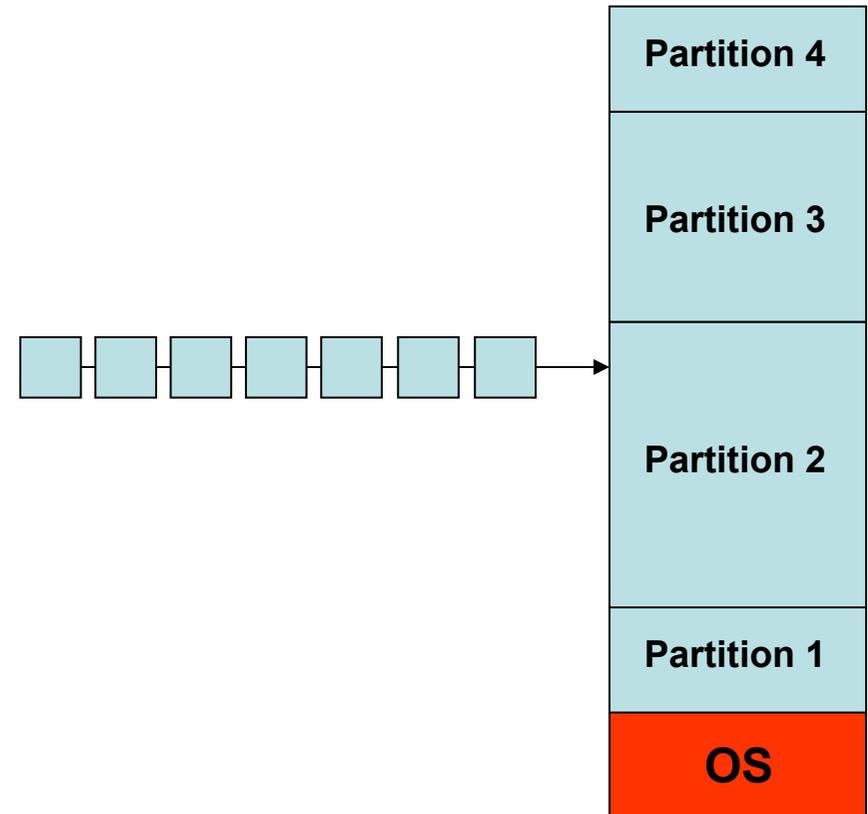
Verwaltung fester Partitionen

mehrere Warteschlangen



Jedes Element einer WS hat spez. Speicherbedarf, der genau mit der zugehörigen Partition übereinstimmt.

eine Warteschlange



Speicherverwaltung mit Partitionen

Relokation und Zugriffsschutz

Relokation:

Probleme: Programme müssen in beliebigen Speicherbereichen funktionieren.

Mechanismen:

1. Statisches Binden von Speicheradressen zur Compilierungszeit ☹️
2. Relokation beim Einlagern in den Hauptspeicher → loader/linker
3. Relokation zur Laufzeit → benötigt positionsunabhängigen Code
→ setzt Unterstützung durch die Rechnerarchitektur voraus.

Zugriffsschutz:

Probleme: Beliebige Referenzen auf Bereiche ausserhalb der Partition.

Mechanismen:

1. Speicherblöcke fester Länge, die mit einem 4-bit Protection-Code gekennzeichnet sind (tagged memory).
Der p-code wird bei jedem Zugriff mit dem entsprechenden Feld des Programm-Status-Worts (PSW) verglichen (IBM 360)
2. Base und Bound Register (CDC 6600)



Relokation

test.c

```
c-program
int main()
{
    exit(0);
}
```

test.s

```
asm-program
main:
    pushl %ebp
    movl %esp,%ebp
    pushl $0
    call exit
    addl $4,%esp
    movl %ebp,%esp
    popl %ebp
    ret
```

test.o

```
link module
0000 55
0001 89E5
0003 6A00
0005 E800000000
000a 83C404
000d 89EC
000f 5D
0010 C3

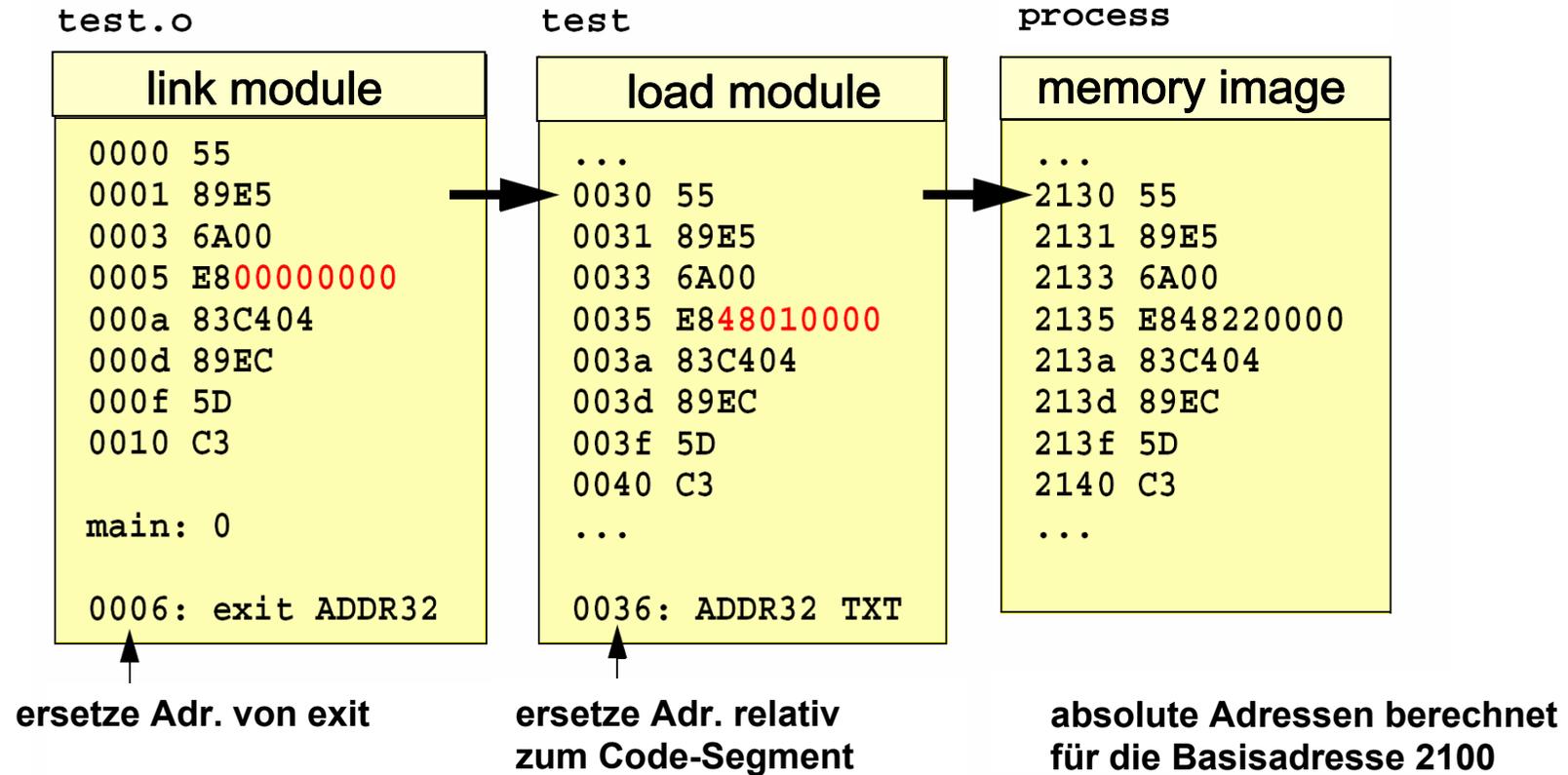
main: 0

0006: exit ADDR32
```

ersetze Adr. von exit

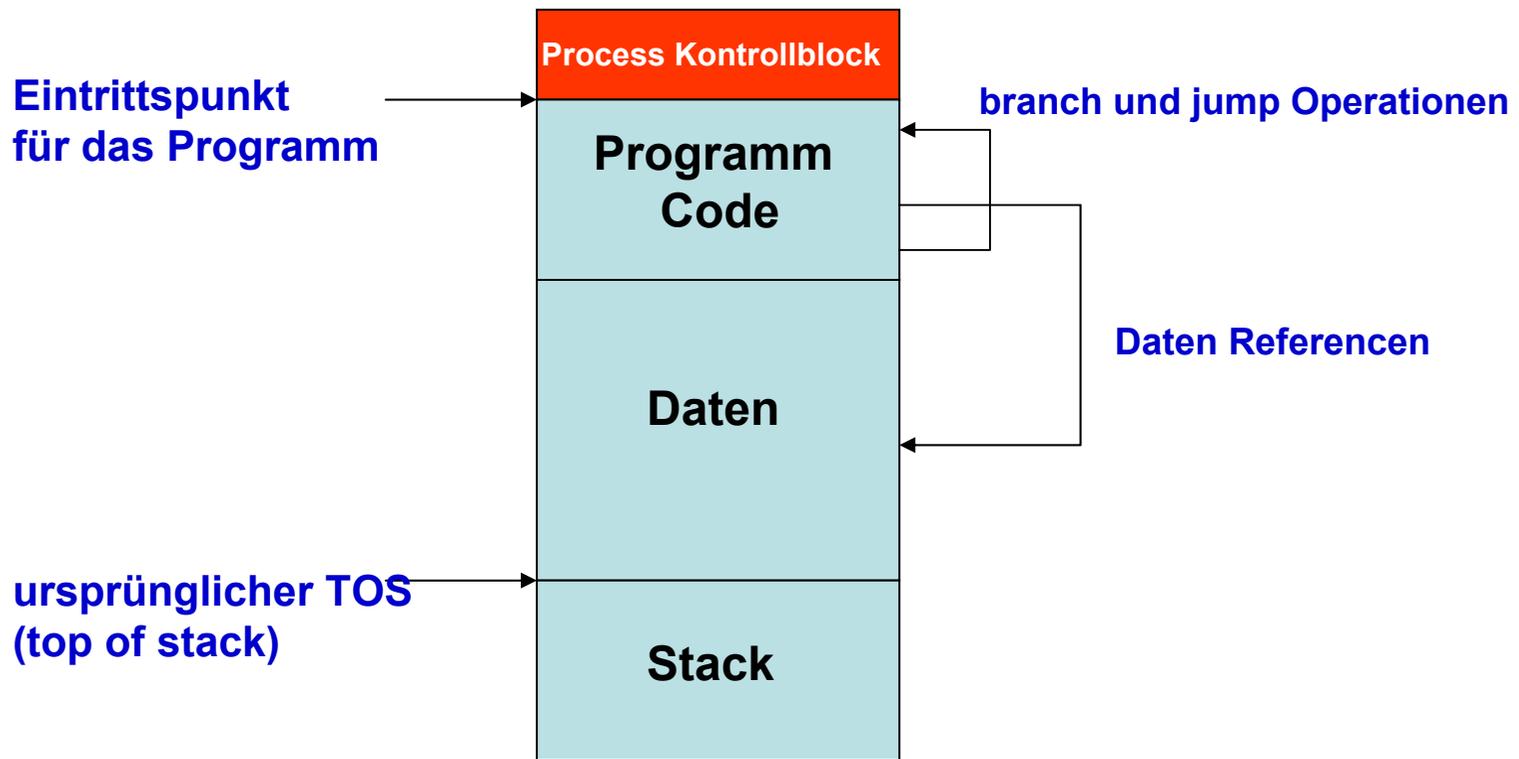


Relokation

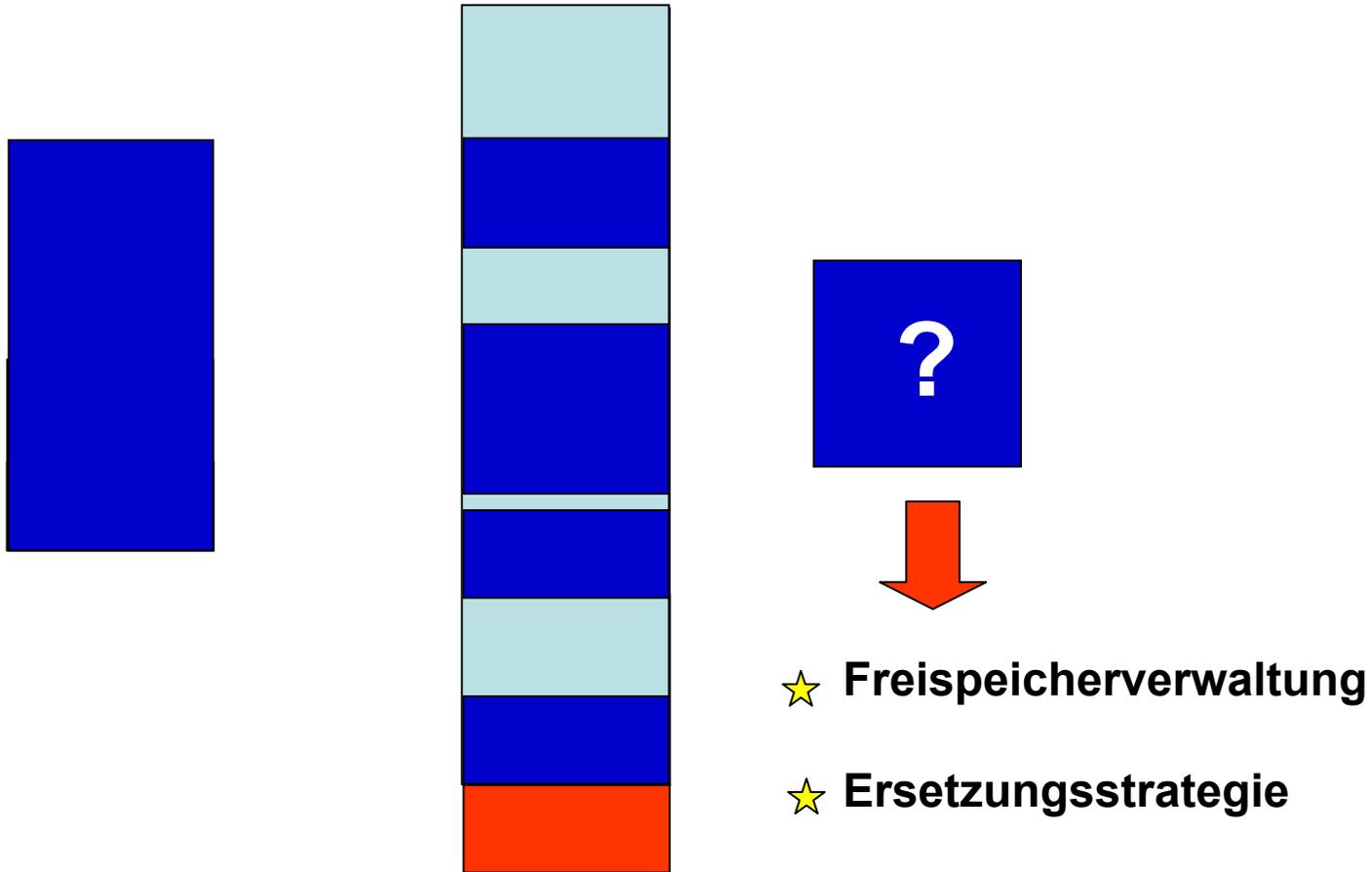


Dynamische Partitionen und Ein-/Auslagern (Swapping)

Ein-/Auslagern: Jeder Prozess wird vollständig vom Speicher auf die Platte geschoben und umgekehrt, mit seinen Code-, Daten- and Stackbereichen.

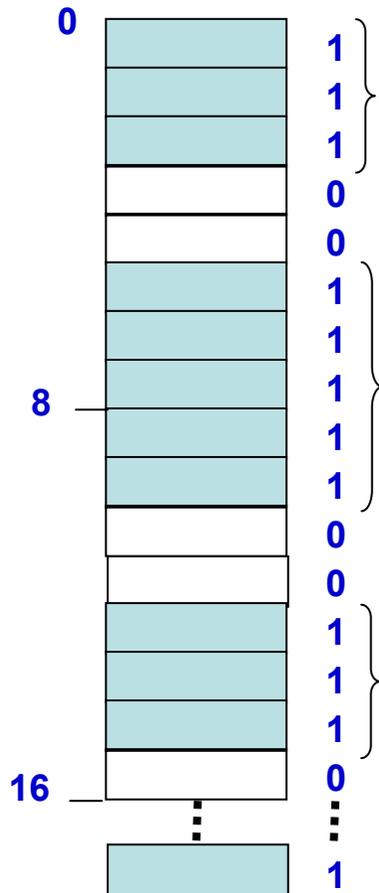


Dynamische Partitionen und Ein-/Auslagern (Swapping)



Dynamische Partitionen und Ein-/Auslagern (Swapping)

Problem: Wo sind freie Blöcke und welche Größe haben sie ?

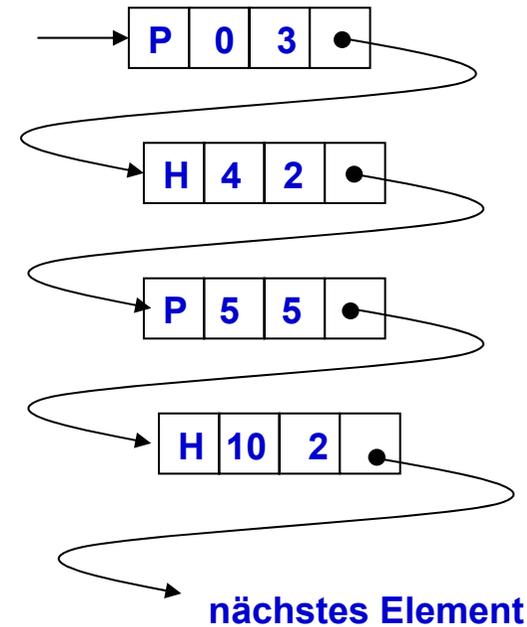


Bitmap-
darstellung:

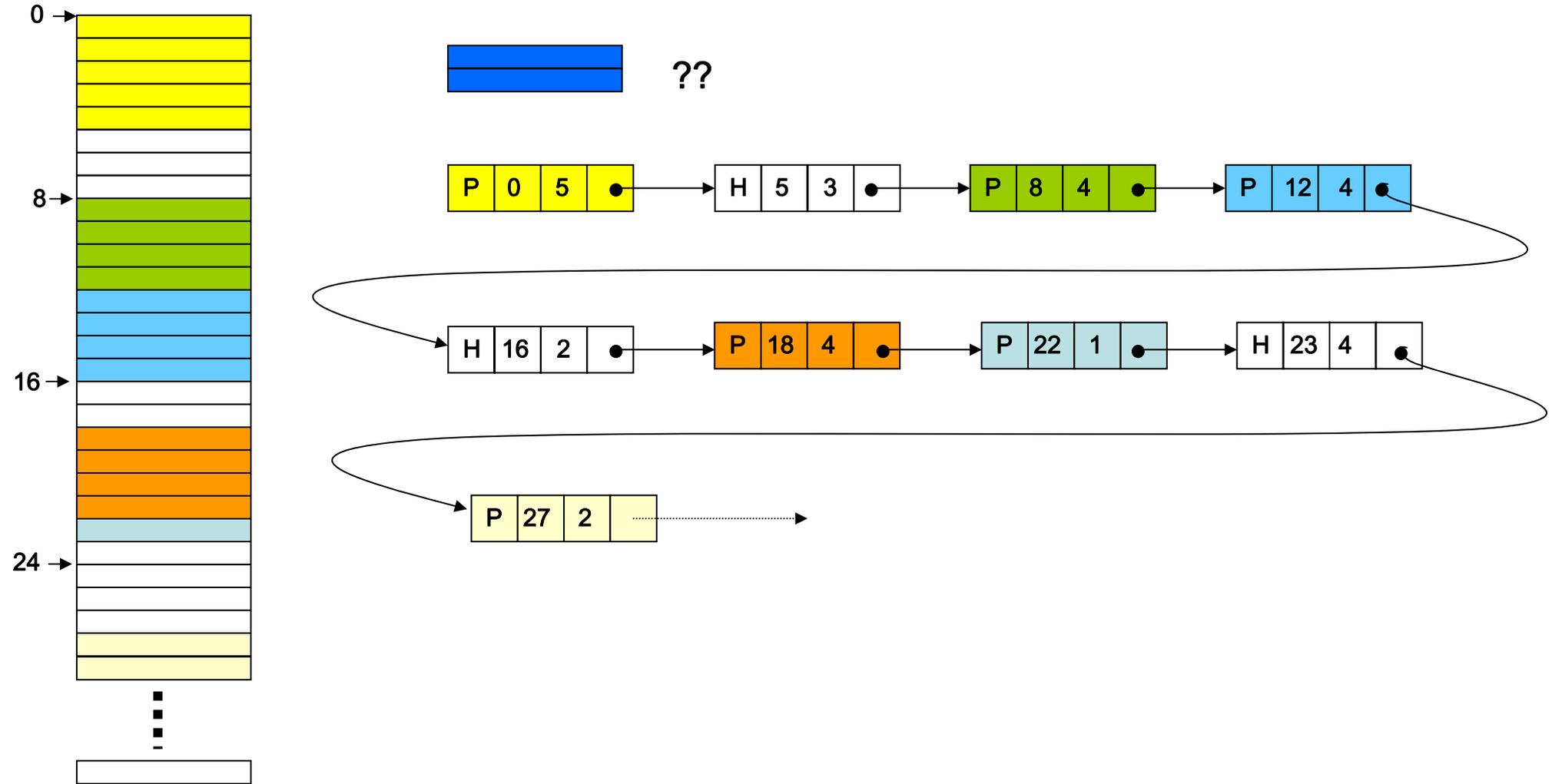
1110011
1110011
10.....

1 = allozierter Block
0 = Loch

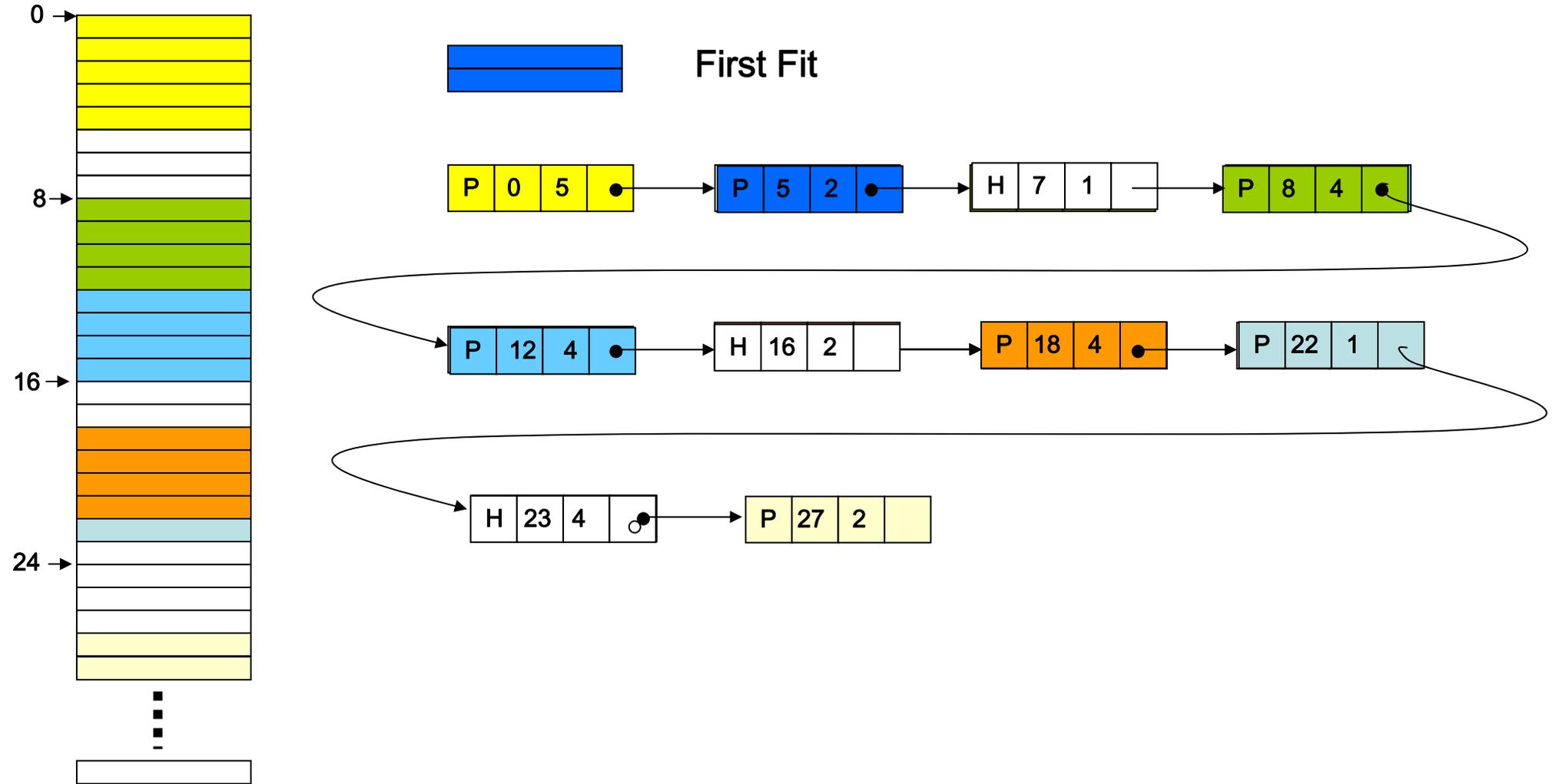
Darstellung durch
verkettete Liste:



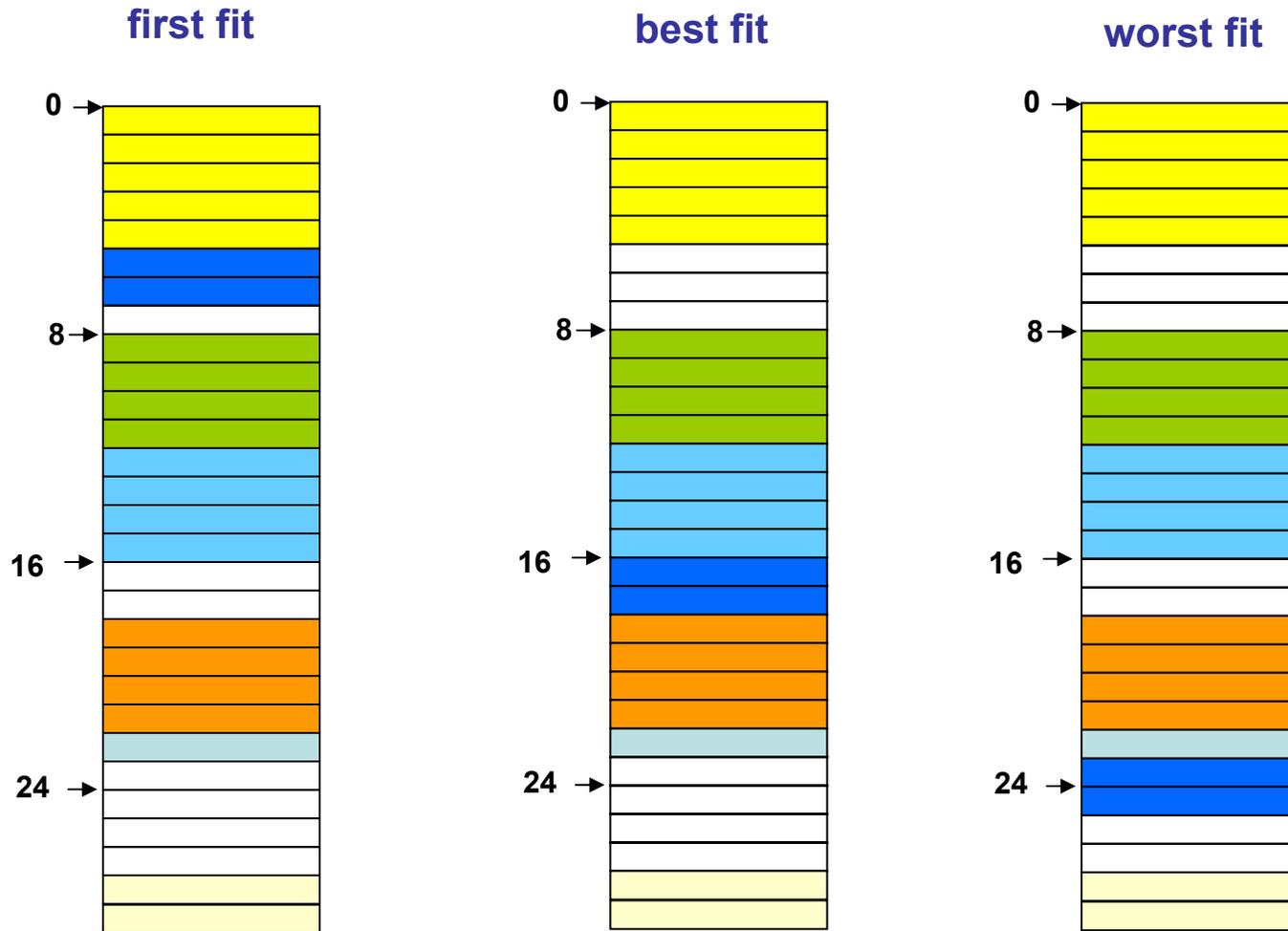
Dynamische Partitionen and Ein-/Auslagern (Swapping)



Dynamische Partitionen and Ein-/Auslagern (Swapping)



Dynamische Partitionen and Ein-/Auslagern (Swapping)

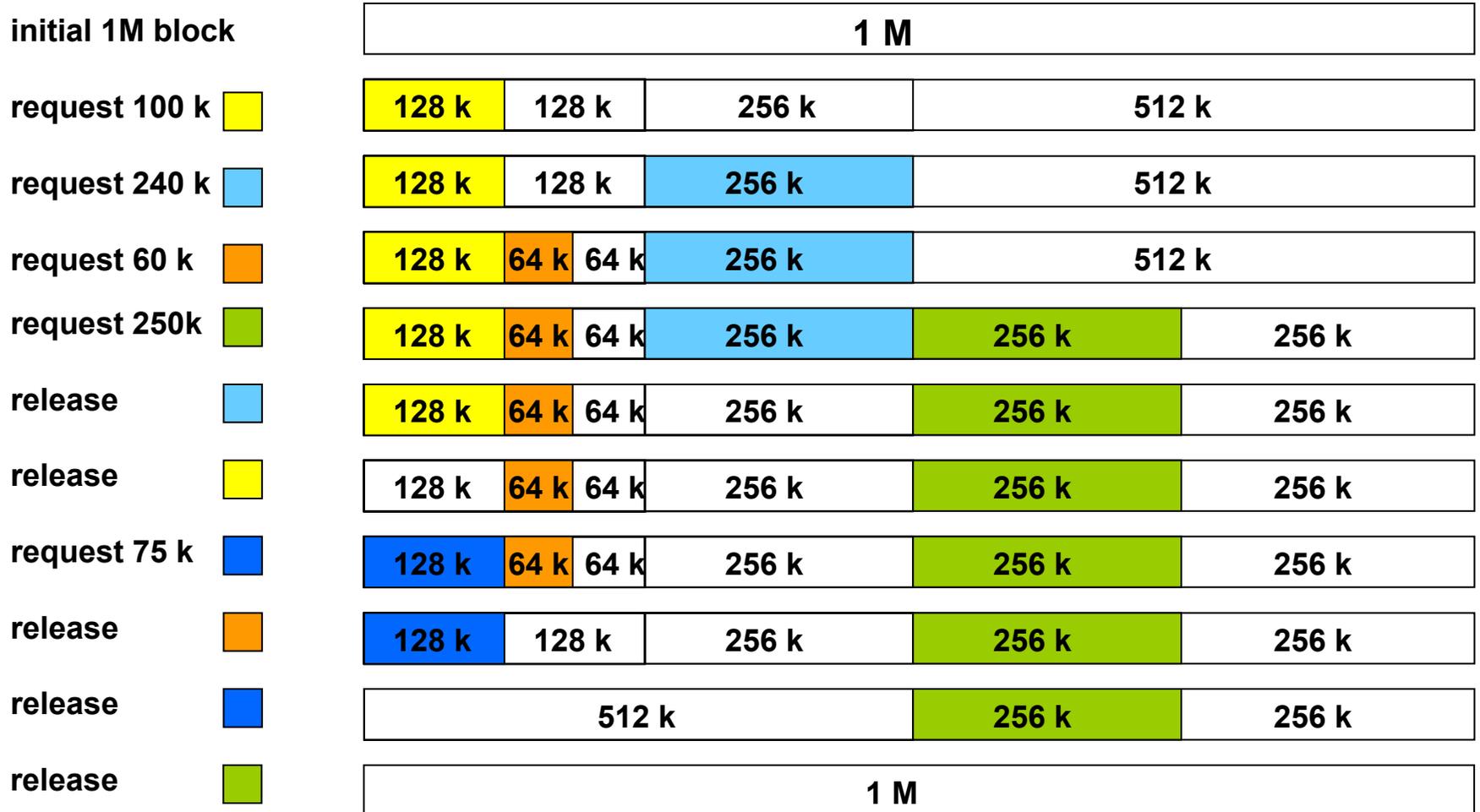


next fit: Startpunkt ist die Position des zuletzt gefundenen Blocks

quick fit: maintains multiple list of free memory blocks according to size of blocks.



Das Buddy-System



Diskussion:

Bisher wurde betrachtet::

- 1. Verwaltung des Realen, physischen Speichers.**
- 2. Der Adressraum ist dem Realspeicher angepasst.**
- 3. Ein- und Auslagern von Speicherblocks wird explizit durch das BS vorgenommen.**
- 4. Die Größe der Einheiten wird durch den Programmierer festgelegt.**

Probleme:

- Programme können größer sein als der Realspeicher.**
- Zugriffsschutz, wenn mehrere Prozesse in einem Adreßraum ablaufen.**



Diskussion:

Adreßraum eines Prozessors:

32 Bit = 4.294.967.296

64 Bit = 18.446.821.383.201.879.616 ~ 2 x 1019**

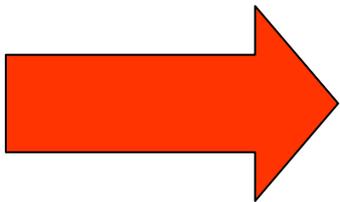


Idee:

**Betrachten des Realspeichers als ein Fenster in einen viel größeren Speicher.
Trennung von logischem und realem Adreßraum.**

Wünschenswerte Ziele:

- 1. Transparenter Mechanismus für das Ein-Aulagern von Speicherblöcken.**
- 2. Logischer Adreßraum ist sehr viel größer als der Adreßraum des Realspeichers.**
- 3. Transparenter Mechanismus zur Relokation.**
- 4. Besserer Zugriffsschutz durch Trennung logischer Adreßräume.**

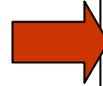


virtueller Speicher



The Computer Journal

Vol. 4, Issue 3,
October 1961



virtual memory also
described in:

John Fotheringham:
Dynamic Storage Allocation
in the ATLAS Including
Automatic Use of Backing
Store

Communications of the ACM,
Volume 4 , Issue 10
(October 1961)

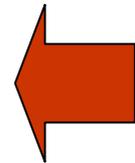
The Manchester University Atlas Operating System

Part I: Internal Organization

By T. Kilburn, D. J. Howarth, R. B. Payne and F. H. Sumner

Introduction

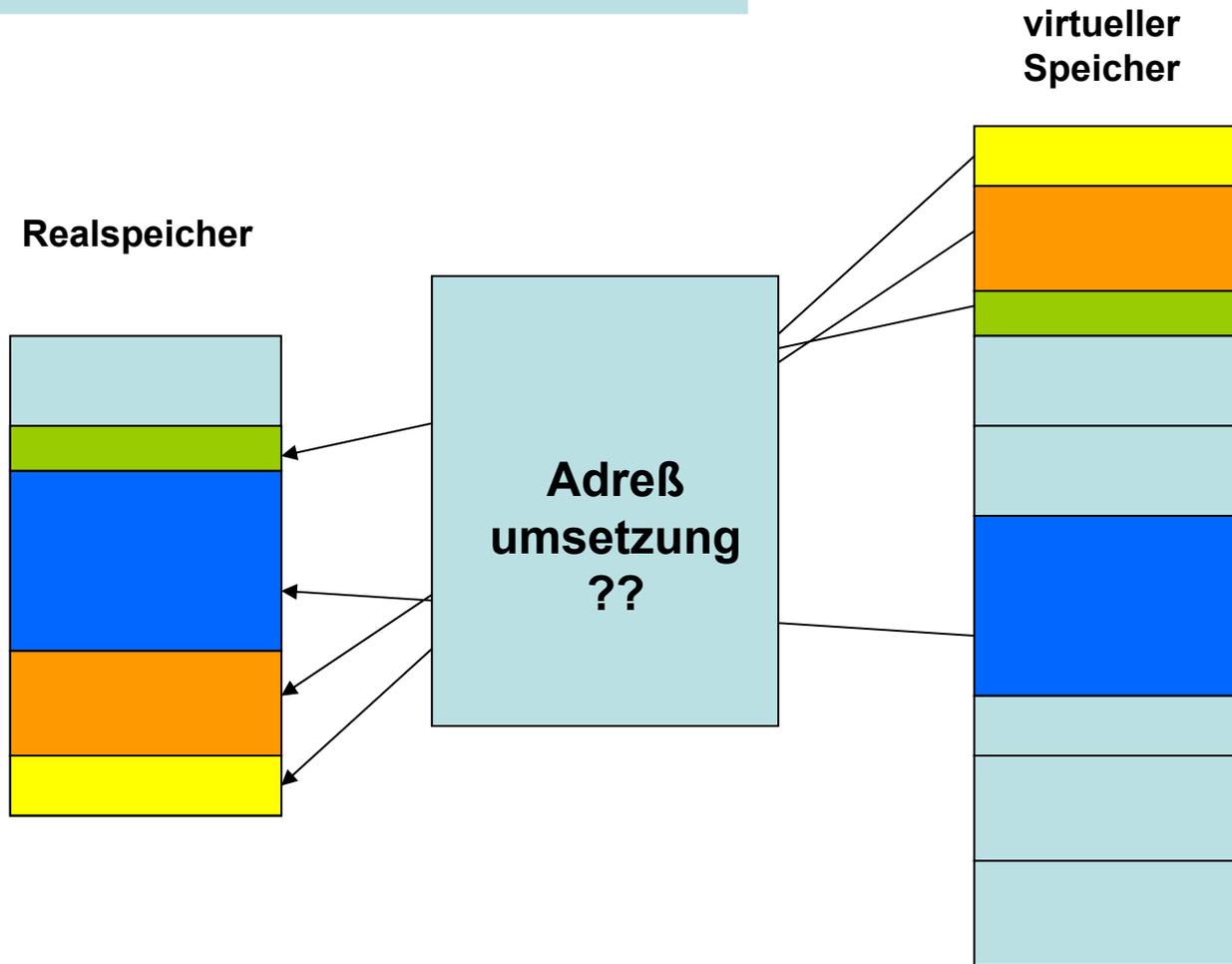
Atlas* is the name given to a comprehensive computer system designed by a joint team of Ferranti Ltd. and Manchester University engineers. The computer system comprises the central computer, fixed store, core store, magnetic drum store, magnetic tapes, and a large quantity and variety of peripheral equipments for input and output. The Manchester University Atlas has 32 blocks of core store each of 512 forty-eight bit words. There is also a magnetic drum store, and transfers between core and drum stores are performed automatically, giving an effective one-level store* of over two hundred blocks. The average time for an instruction is between 1 and 2 microseconds. The peripheral equipments available on the Manchester University Atlas include



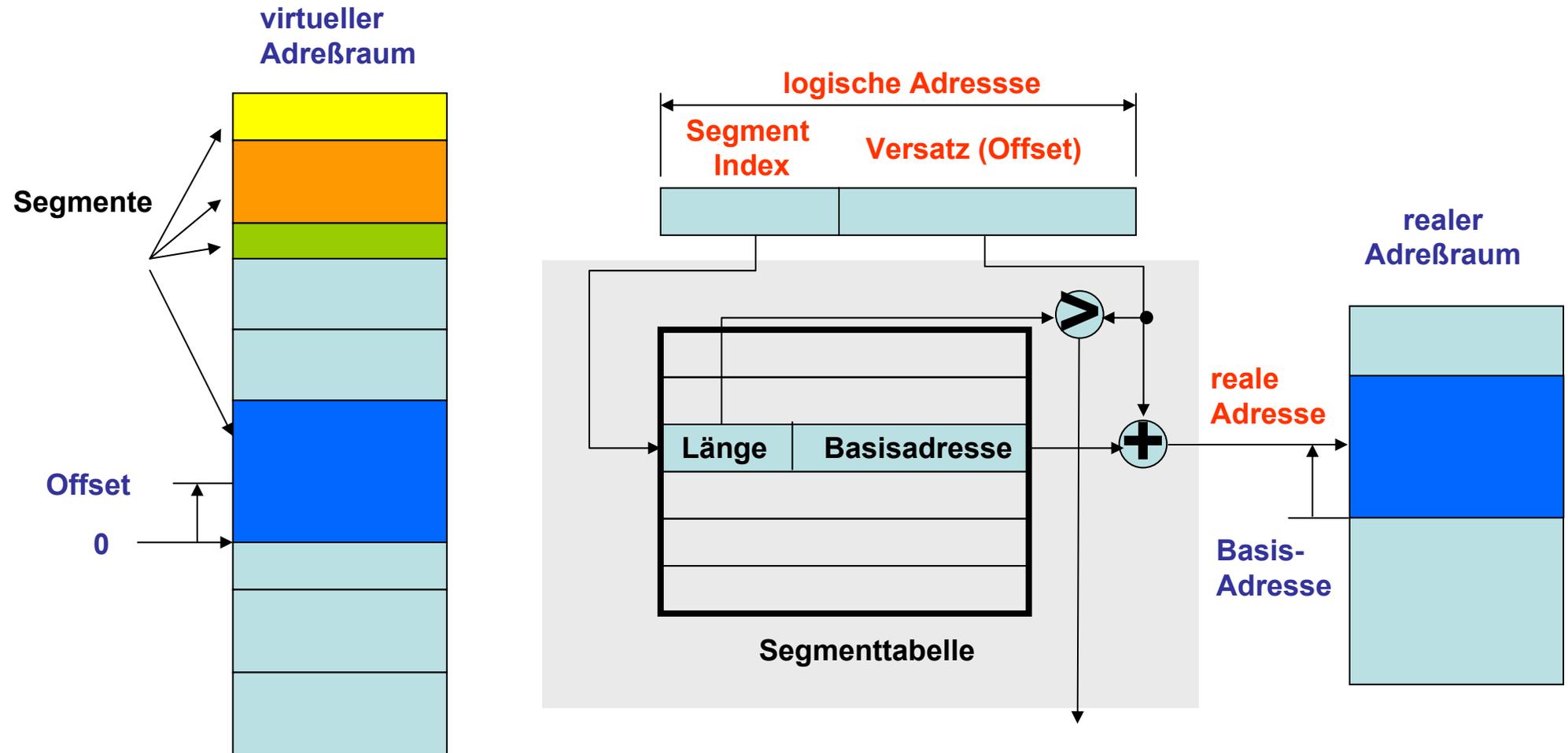
8 magnetic tape decks	90,000 characters per second
4 paper tape readers	300 characters per second
4 paper tape punches	110 characters per second
1 line printer	600 lines per minute
1 card reader	600 cards per minute
1 card punch	100 cards per minute



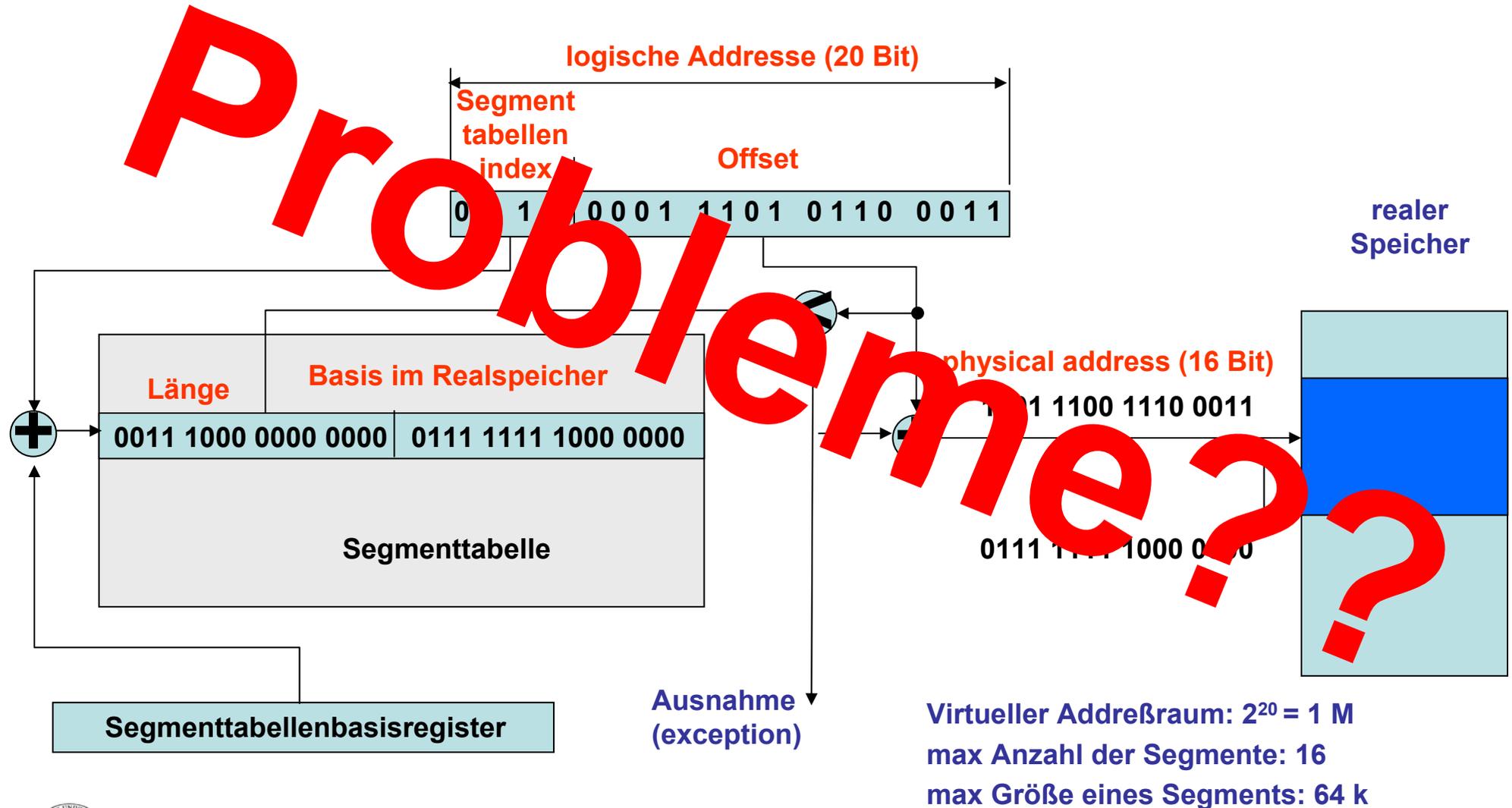
virtual memory



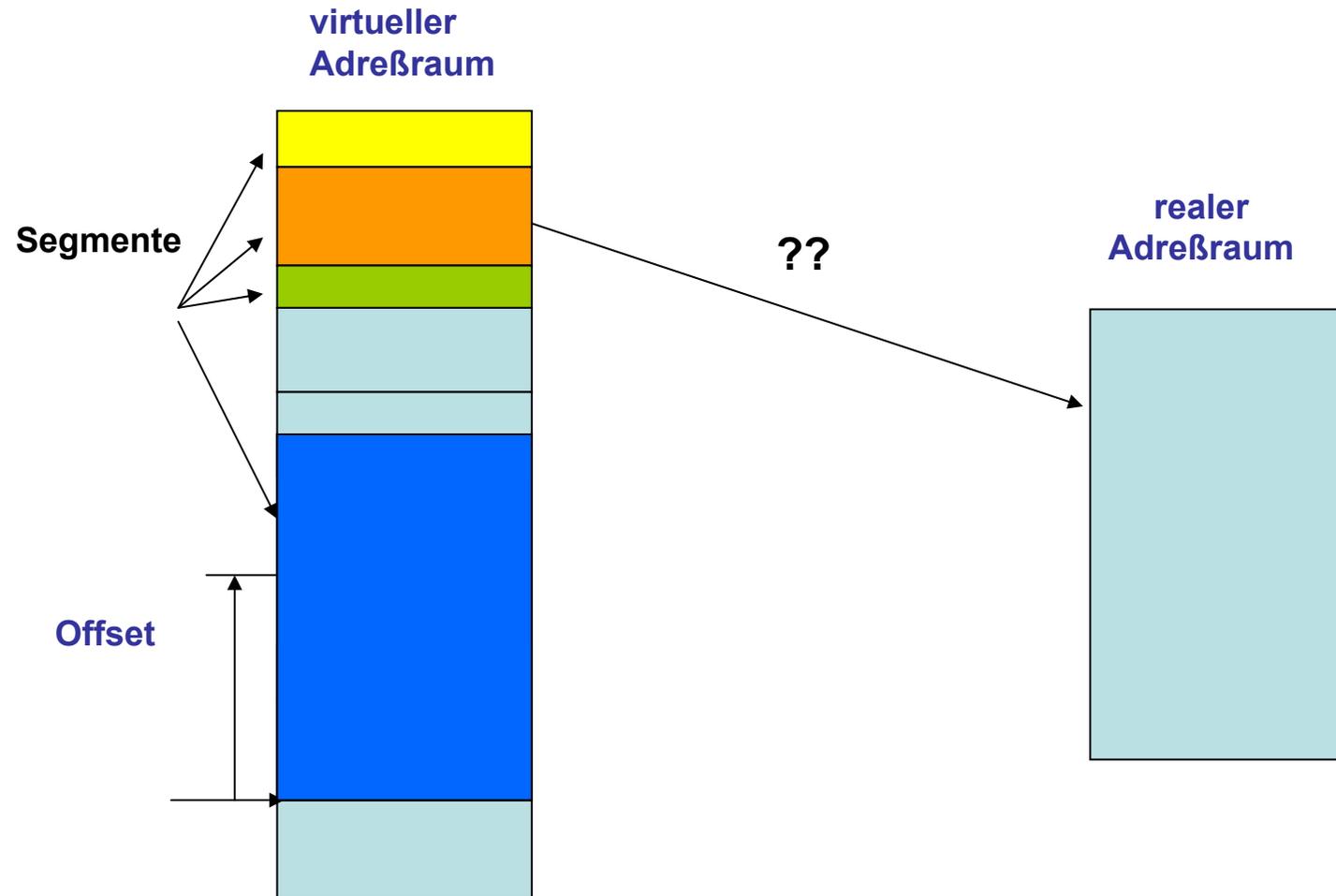
"segmentierter" virtueller Speicher



Segmentierter Speicher - Adreßumsetzung



"segmentierter" virtueller Speicher



Seitenorientierter virtueller Speicher

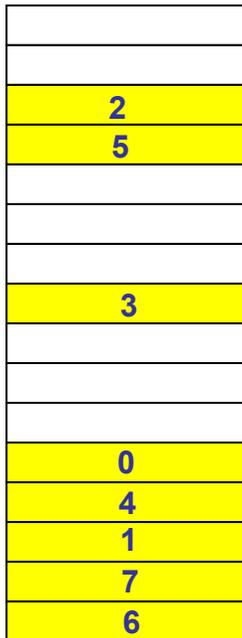


Idee:

Seiten fester Größe werden auf Kacheln (Seitenrahmen: frames) fester Größe im Hauptspeicher abgebildet.

virtueller
Adreßraum

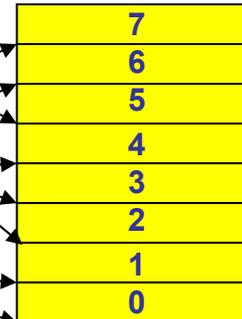
60k - 64k
56k - 60k
52k - 56k
48k - 52k
44k - 48k
40k - 44k
36k - 40k
32k - 36k
28k - 32k
24k - 28k
20k - 24k
16k - 20k
12k - 16k
8k - 12k
4k - 8k
0k - 4k



(virtuelle)
Seiten

Realspeicher-
Adresse

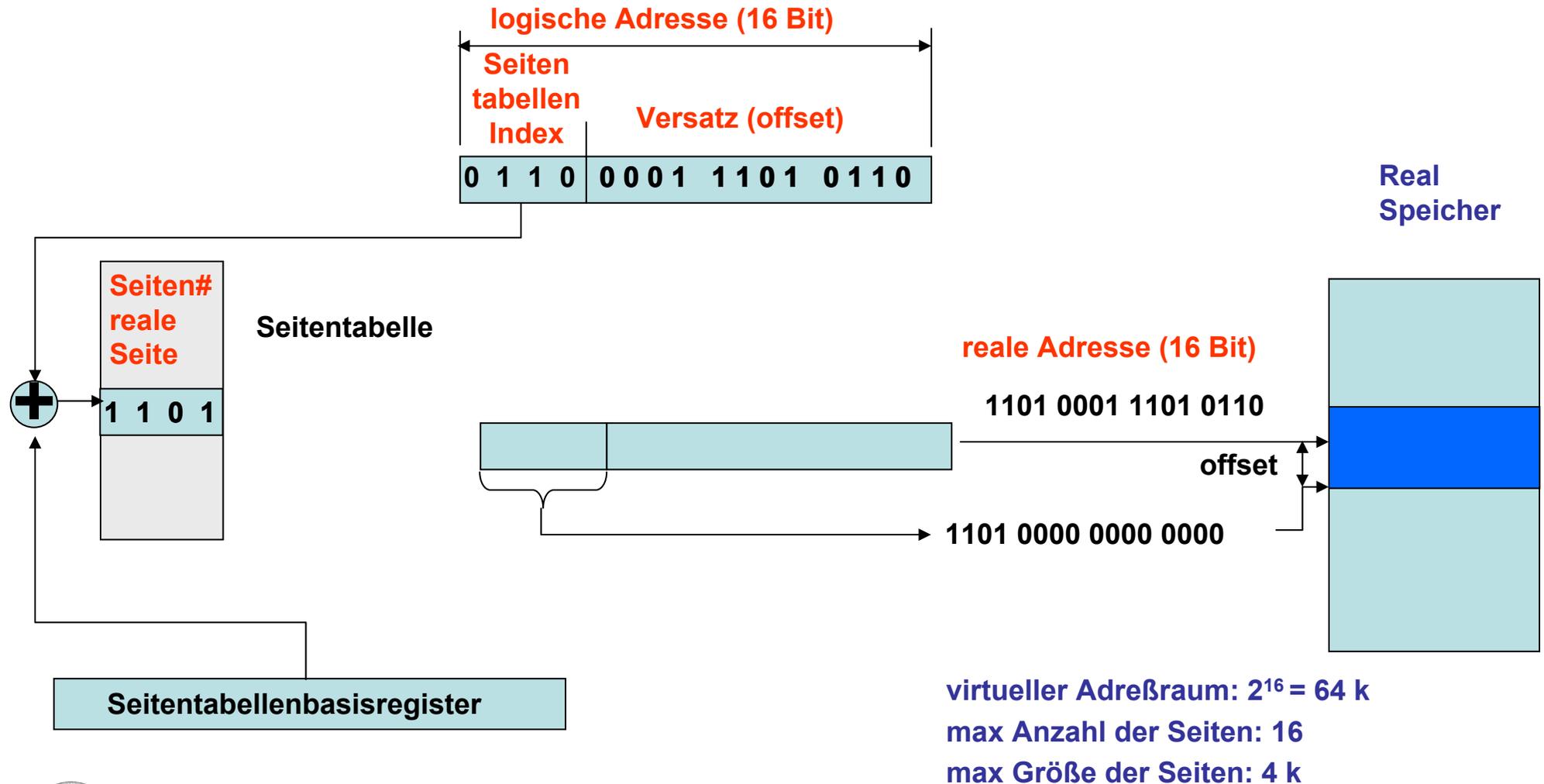
28k - 32k
24k - 28k
20k - 24k
16k - 20k
12k - 16k
8k - 12k
4k - 8k
0k - 4k



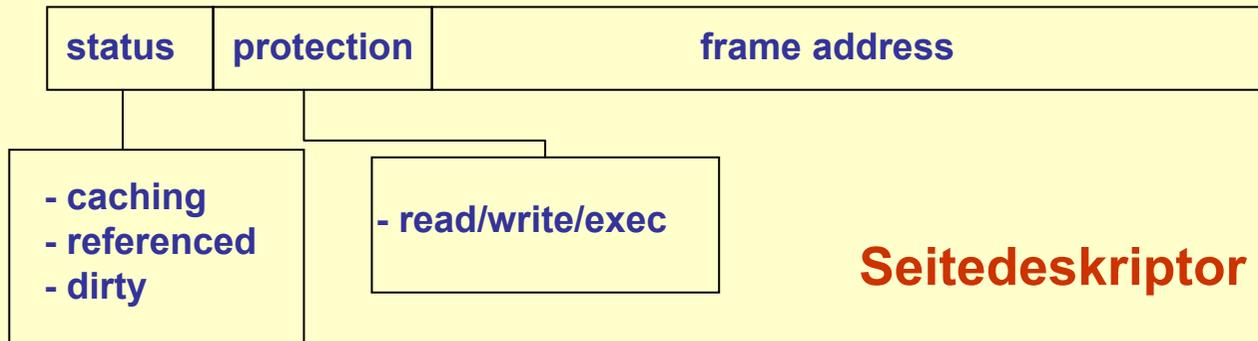
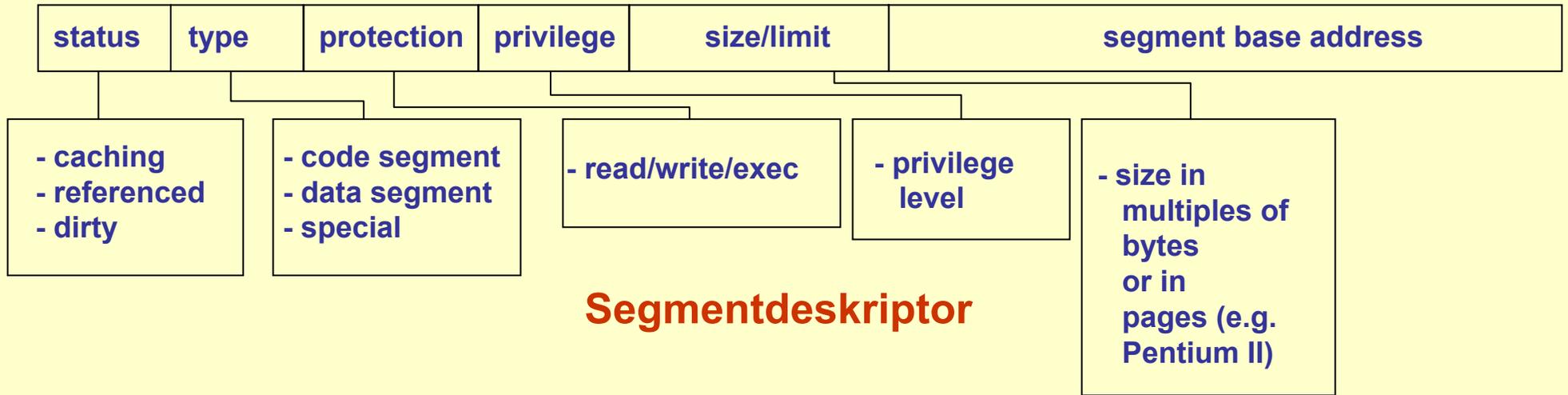
(reale)
Seitenrahmen



Adreßumsetzung mit Seiten



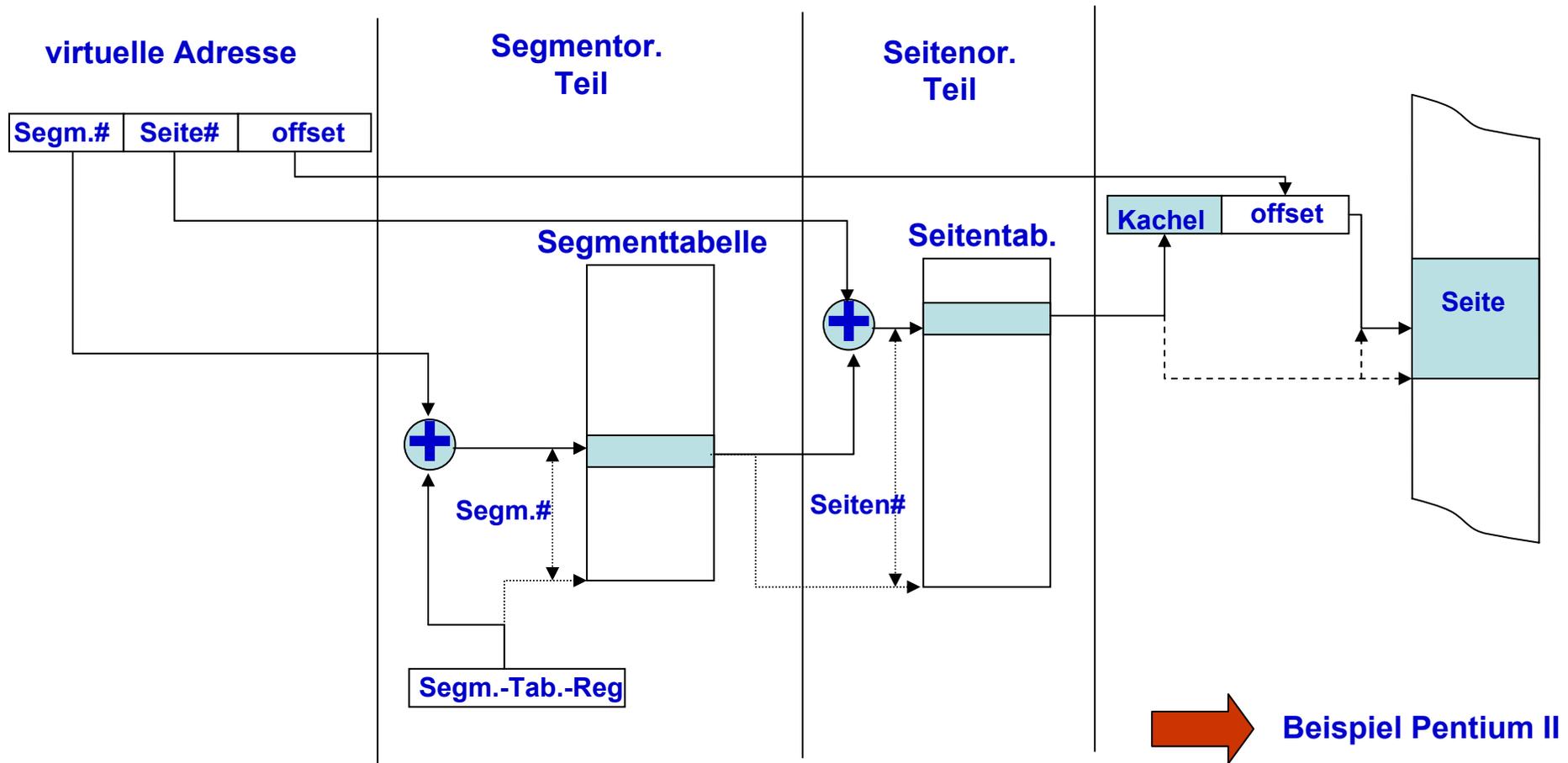
Struktur eines Tabelleneintrags



Diskussion: Segmente gegen Seiten

	Transparenz für Progr.	Anzahl d. Adreßräume	virt. Adr. raum > realer Speicher	variable Objektgr.	Fragment.	Verw. overhead	Hauptgrund für Einführung
Seitenor.		ein			intern		unendlich viel Speicher
Segmentor.		viele			extern		mehrere Adr. räume

Kombination von Segmenten und Seiten



Größe der Seitentabellen

32-Bit Adreßraum: 4G Addresses

Seitentabellengröße @ 4k: 1M @ 4k entries

64-Bit Adreßraum: 4G·4G Addresses

Seitentabellengröße @ 4k: 4G·1M entries

1. Seitengröße erhöhen: e.g. UltraSPARC II supports 8k, 64k, 512k and 4M pages

Grund: weniger Seiten

Problem: interne Fragmentierung

2. Seitentabellen ein-und auslagern: mehrstufige Seitentabelle

Grund: a.) Virtueller Teil Speicher ist billig, b.) es wird aktuell immer nur ein kleiner der Seiten benötigt (→ working set)

Problem: eine zusätzliche Indirektionsstufe

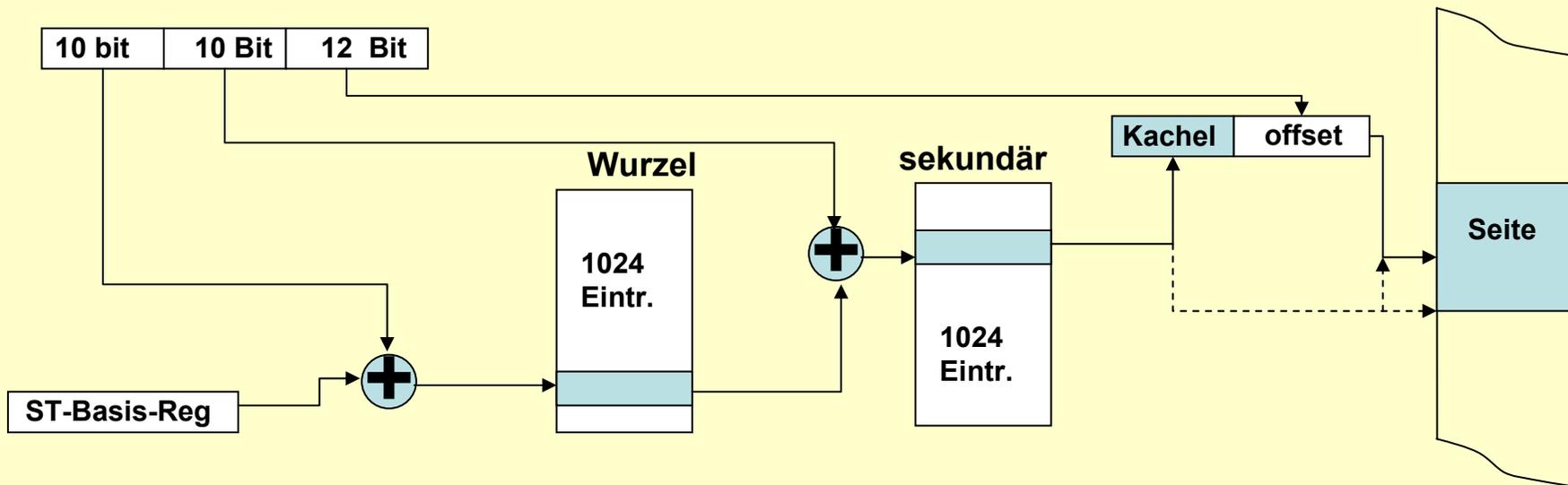
3. Umsetzen von realen auf virtuelle Adressen: Invertierte Seitentabellen

Grund: Der reale Adreßraum ist klein im Vergleich zum virtuellen.

Problem: viel mehr Seiten als Seitenrahmen → keine eindeutige Abbildung



mehrstufige hierarchische Seitentabellen



invertiert oder nicht invertiert?

Pro invertiert:

Seitentabelle ist proportional zur Größe des Realspeichers:

z.B. für 1 G @ 8k Seiten: 128k Einträge unabhängig von Größe d. virt. A-Raums

Con invertiert:

- Hashing muss bei jedem Speicherzugriff durchgeführt werden.
- Verkettung kann zu mehrfachen Zugriff führen.
- benötigt komplexe Ersetzungs-und Verwaltungsstrategien.

ABER:

- Hierarchische Seitentabellen benötigen ebenfalls mehrfachen Zugriff.
- wenn sekundäre Seitentabellen ausgelagert sind: mehrfacher Plattenzugriff.

 **Hardwareunterstützung wird in allen Fällen benötigt !**

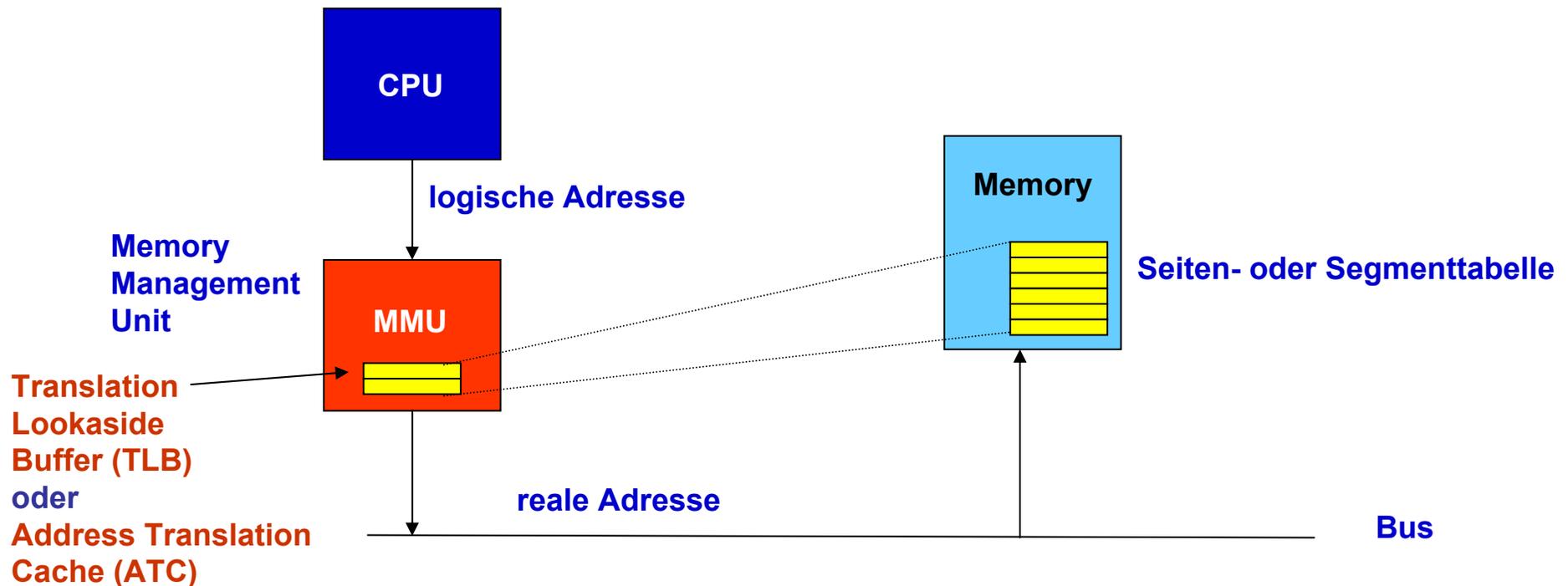


Die MMU: Speicherverwaltungseinheit

Adreßübersetzung benötigt mehrere Ebenen der Indirektion

➔ negative Auswirkung auf Leistung !

➔ benötigt Hardwareunterstützung zur Beschleunigung.



Verwaltung des virtuellen Speichers

Woher weiss man, wie viele Seiten benötigt werden?

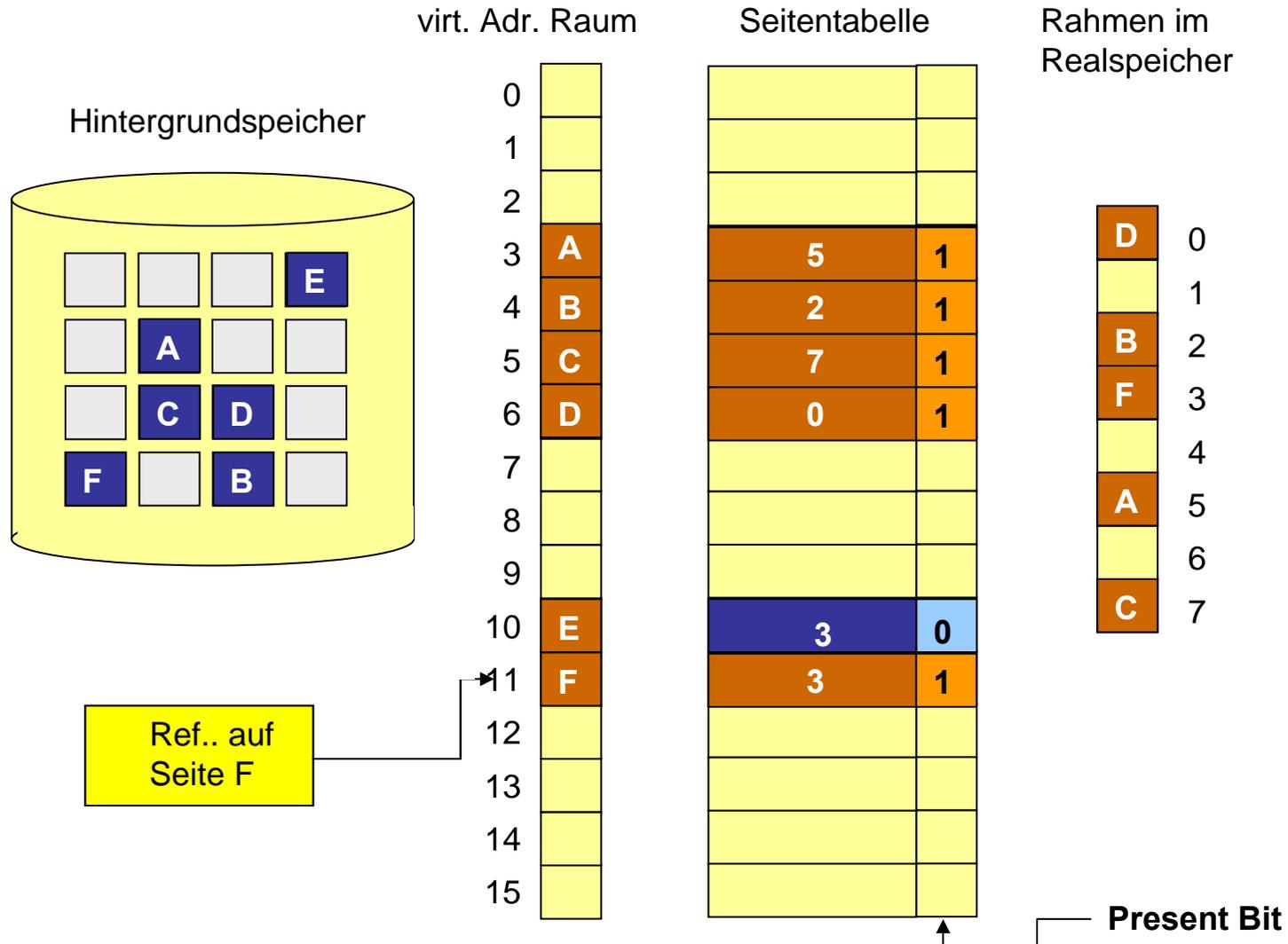
➔ Einlagern von Seiten auf Anforderung

Was tun, wenn mehr Seiten benötigt werden als Seitenrahmen zur Verfügung stehen?

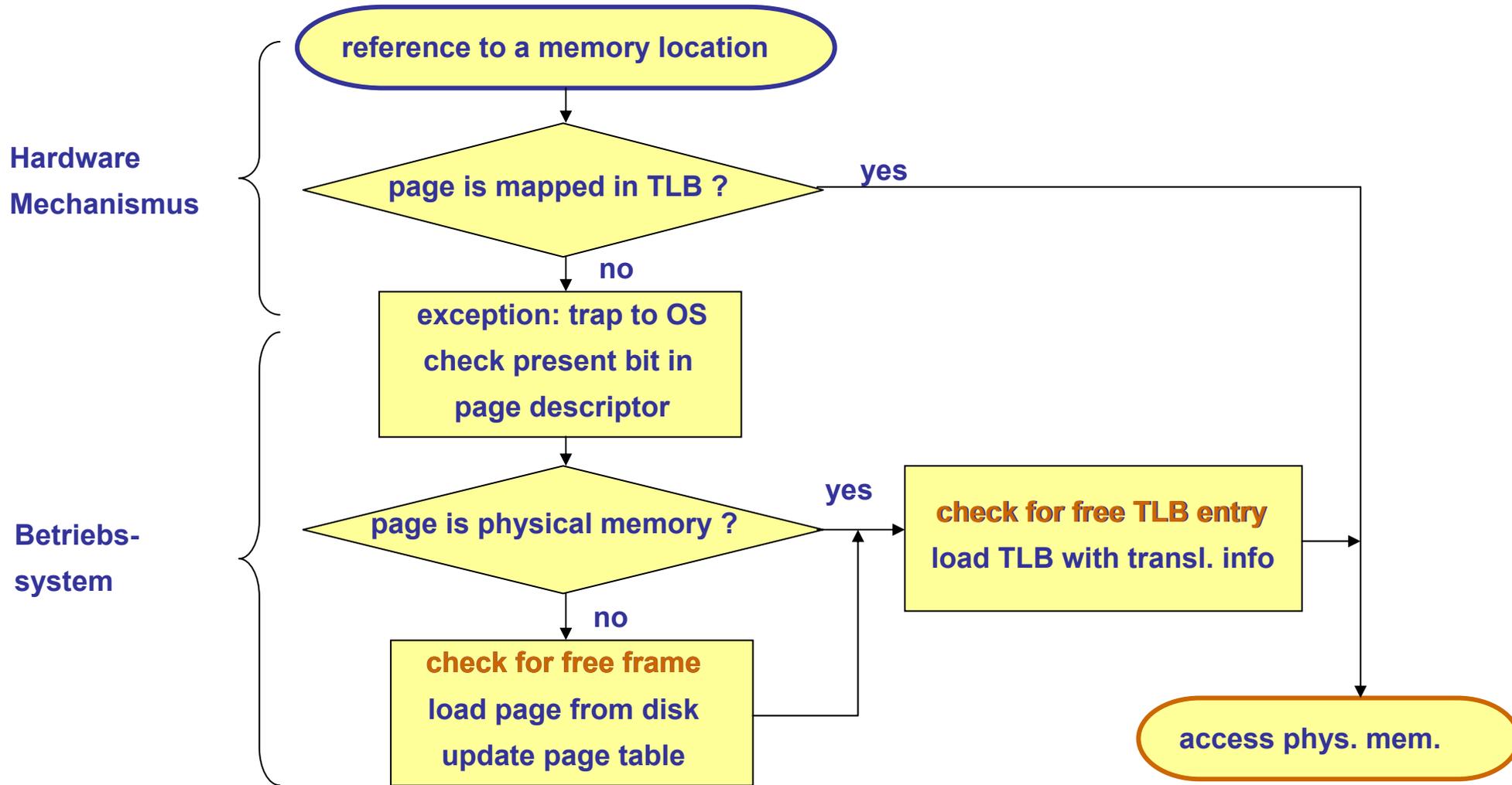
➔ Seitenersetzungsmechanismen



Demand Paging



Demand Paging



Demand Paging: Diskussion

Normaler Zugriff, kein Seitenfehler: Zugriffszeit ~ 5 - 200 n

Wie hoch ist die Zugriffszeit bei einem Seitenfehler?

Annahmen {
p: Wahrscheinlichkeit für einen Seitenfehler
Normale effektive Zugriffszeit: 100 ns
Laden einer Seite von der Platte: ~ 20 ms

Effektive Zugriffszeit für $p=0,01$:

$$(1-p) \cdot 100 + p \cdot 20.000.000 = 0,99 \cdot 100 + 0,01 \cdot 20.000.000 = 99 \cdot 200.000 \text{ ns} = 198 \mu\text{s}$$

Um im Bereich der normalen Zugriffszeit zum Realspeicher zu bleiben muss die Wahrscheinlichkeit eines Seitenfehlers in der Größenordnung von 0,000005 liegen !

➔ 1 Seite pro 200000 Zugriffe darf zu einem Seitenfehler führen!



Seitenersetzungsverfahren

Das Betriebssystem muss zukünftige Anforderungen vorhersagen. Dazu kann es nur die Analyse der Gegenwart und Vergangenheit ausnutzen.

- Wurde auf der Seite zugegriffen?
- Wurde die Seite modifiziert?
- Welche Prozesse sind zur Zeit aktiv?



Seiten-
Deskriptor



C: Caching, **R:** Referenced, **D:** Dirty (modified), **P:** Present



Optimale Seitenersetzungsstrategien

ref. sequence		1	2	3	4	1	2	5	1	2	3	4	5
frame assignment in phys. memory	frame 1	1	1	1	1	1	1	1	1	1	3	4	4
	frame 2		2	2	2	2	2	2	2	2	2	2	2
	frame 3			3	4	4	4	5	5	5	5	5	5
control state: distance to next reference	frame 1	4	3	2	1	3	2	1	∞	∞	∞	∞	∞
	frame 2	∞	4	3	2	1	3	2	1	∞	∞	∞	∞
	frame 3	∞	∞	7	7	6	5	5	4	3	2	2	∞
		P	P	P	P			P			P	P	

3 Kacheln

7 Seitenfehler

ref. sequence		1	2	3	4	1	2	5	1	2	3	4	5
frame assignment in phys. memory	frame 1	1	1	1	1	1	1	1	1	1	1	4	4
	frame 2		2	2	2	2	2	2	2	2	2	2	2
	frame 3			3	3	3	3	3	3	3	3	3	3
	frame 4				4	4	4	5	5	5	5	5	5
control state: distance to next reference	frame 1	4	3	2	1	3	2	1	∞	∞	∞	∞	∞
	frame 2	∞	4	3	2	1	3	2	1	∞	∞	∞	∞
	frame 3	∞	∞	7	6	5	4	3	2	1	∞	∞	∞
	frame 4	∞	∞	∞	7	6	5	5	4	3	2	1	∞
		P	P	P	P	-	-	P	-	-	-	P	-

4 frames

6 page faults

Seitenersetzungsstrategien

Not-recently-used → unterscheidet 4 Seitenklassen:

Klasse 0: R=0, D=0

Klasse 1: R=0, D=1

Klasse 2: R=1, D=0

Klasse 3: R=1, D=1

Problem

NRU ersetzt eine beliebige Seite aus der niedrigsten nichtleeren Klasse

Ref. Folge		sweep				sweep							
		1	2	3	4	1	2	5	1	2	3	4	5
Kachelzuordnung im Realspeicher	frame 1	1	1	1	1	1	1	5	1	1	3	3	5
	frame 2		2	2	2	2	2	2	2	2	2	2	5
	frame 3			3	4	4	4	4	4	4	4	4	4
Kontrollstatus: Seitenklasse	frame 1	2	3	3	3	1	1	2	2	0	2	2	2
	frame 2	-	2	2	2	0	2	3	3	1	1	1	2
	frame 3	-	-	2	2	0	2	3	3	1	3	2	2
		P	P	P	P			P	P		P		P

8 page faults



Seitenersetzungsstrategien

FIFO: Ersetzt die Seite, die am längsten im Speicher ist.

Ref. Folge		1	2	3	4	1	2	5	1	2	3	4	5
Kachelzuordnung im Realspeicher	frame 1	1	1	1	4	4	4	5	5	5	5	5	5
	frame 2		2	2	2	1	1	1	1	1	3	3	3
	frame 3			3	3	3	2	2	2	2	2	4	4
Kontrollstatus: Alter der Seite	frame 1	0	1	2	0	1	2	0	1	2	3	4	5
	frame 2	-	0	1	2	0	1	2	3	4	0	1	2
	frame 3	-	-	0	1	2	0	1	2	3	4	0	1
		P	P	P	P	P	P	P			P	P	

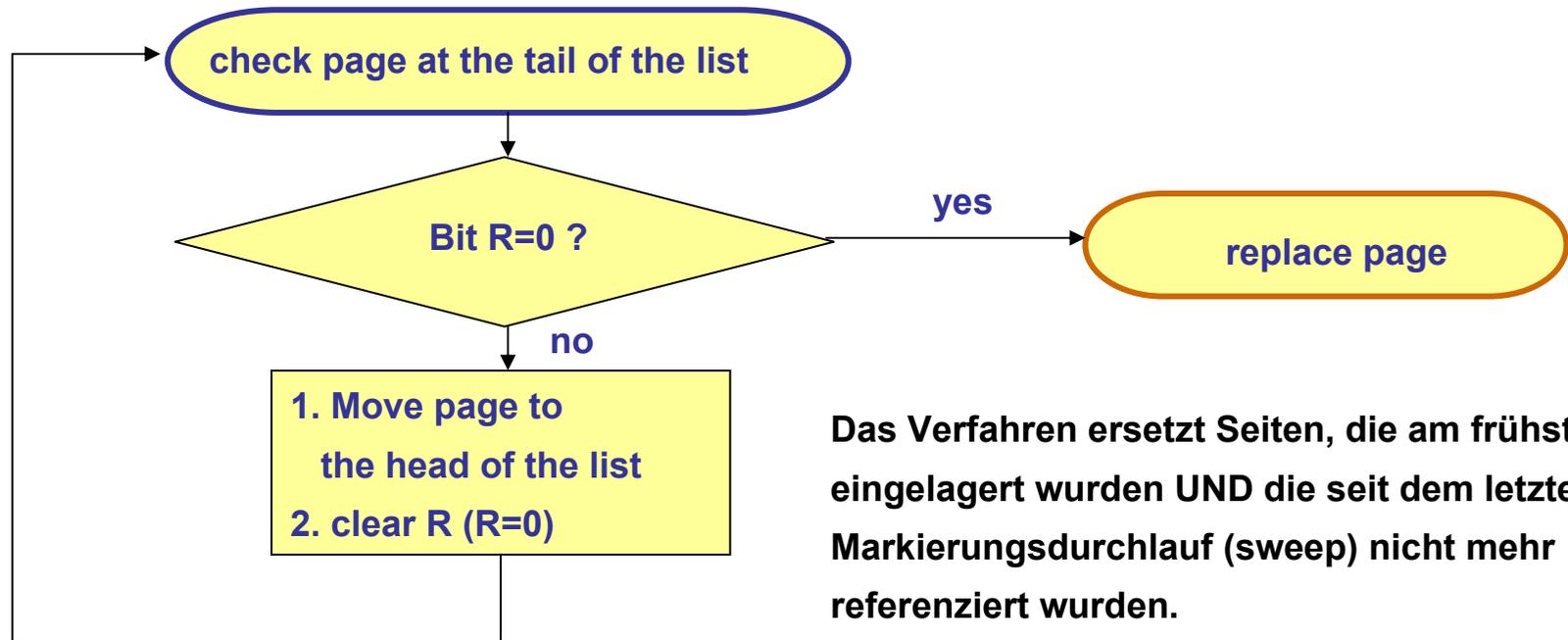
9 Seitenfehler



Seitenersetzungsstrategien

Variation von FIFO: Der "Second Chance" Algorithmus

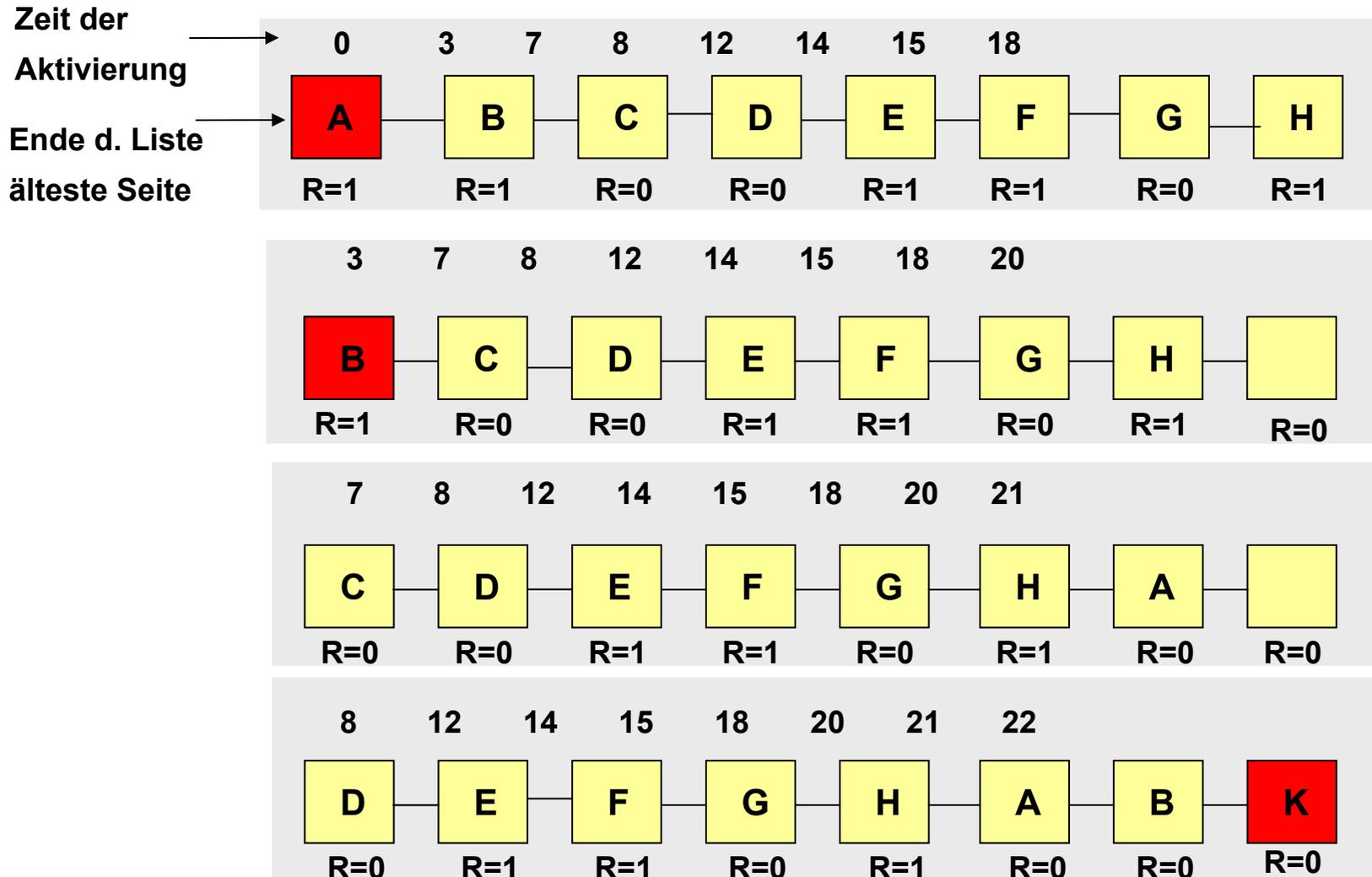
Die Seiten sind in einer Liste nach FIFO geordnet



Das Verfahren ersetzt Seiten, die am frühesten eingelagert wurden UND die seit dem letzten Markierungsdurchlauf (sweep) nicht mehr referenziert wurden.



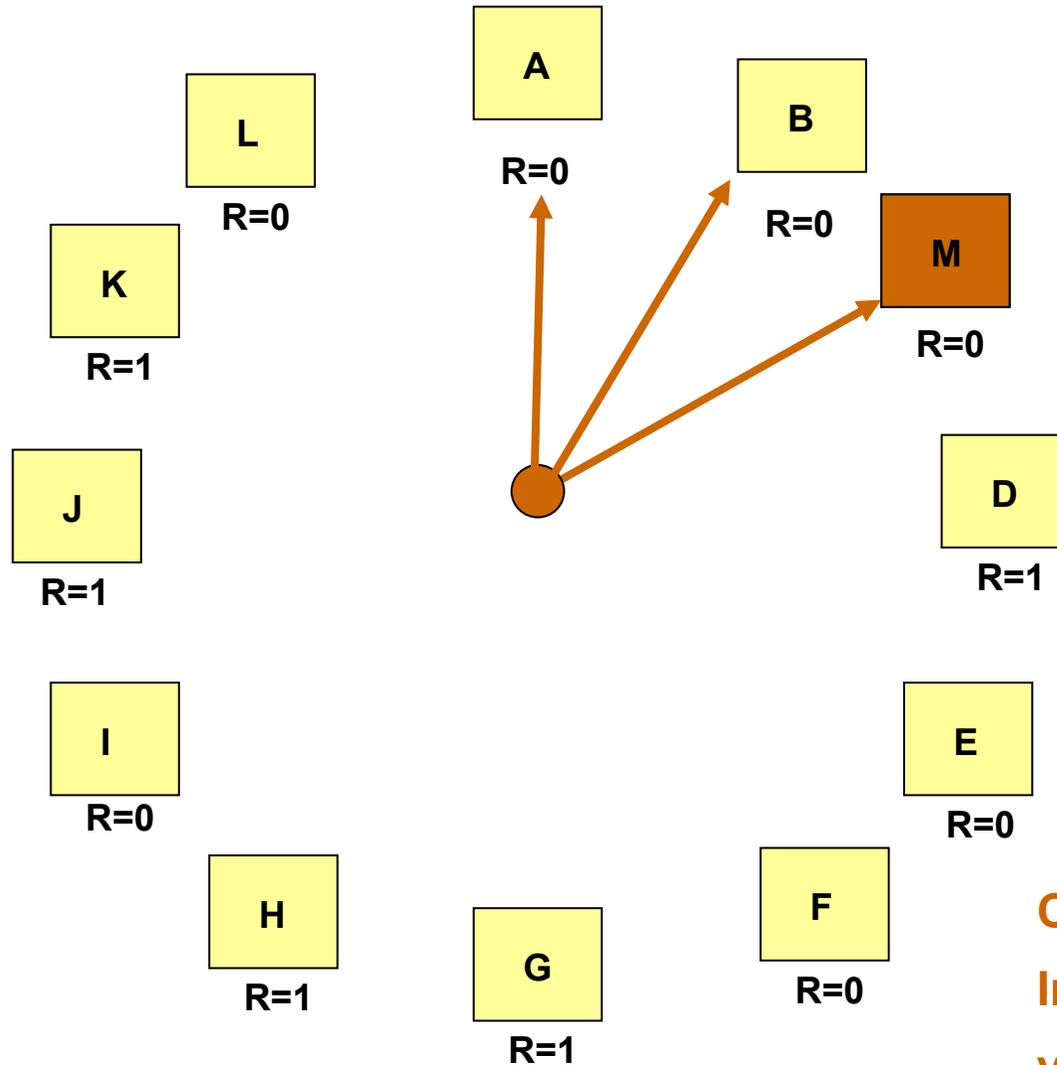
Der "Second Chance" Algorithmus



Seiten sind nach FIFO geordnet



Implementierung von "second chance": Der "Clock" Algorithmus



Clock ist eine intelligente Implementierung von "Second Chance"



Seitenersetzungsstrategien

FIFO: Belady's Anomalie

Ref. Folge		1	2	3	4	1	2	5	1	2	3	4	5
Kachelzuordnung im Realspeicher	frame 1	1	1	1	1	1	1	5	5	5	5	4	4
	frame 2		2	2	2	2	2	2	1	1	1	1	5
	frame 3			3	3	3	3	3	3	2	2	2	2
	frame 4				4	4	4	4	4	4	3	3	3
Kontrollstatus: Alter der Seite	frame 1	0	1	2	3	4	5	0	1	2	3	0	1
	frame 2	-	0	1	2	3	4	5	0	1	2	3	0
	frame 3	-	-	0	1	2	3	4	5	0	1	2	3
	frame 4	-	-	-	0	1	2	3	4	5	0	1	2

P P P P - - P P P P P P 10 Seitenfehler

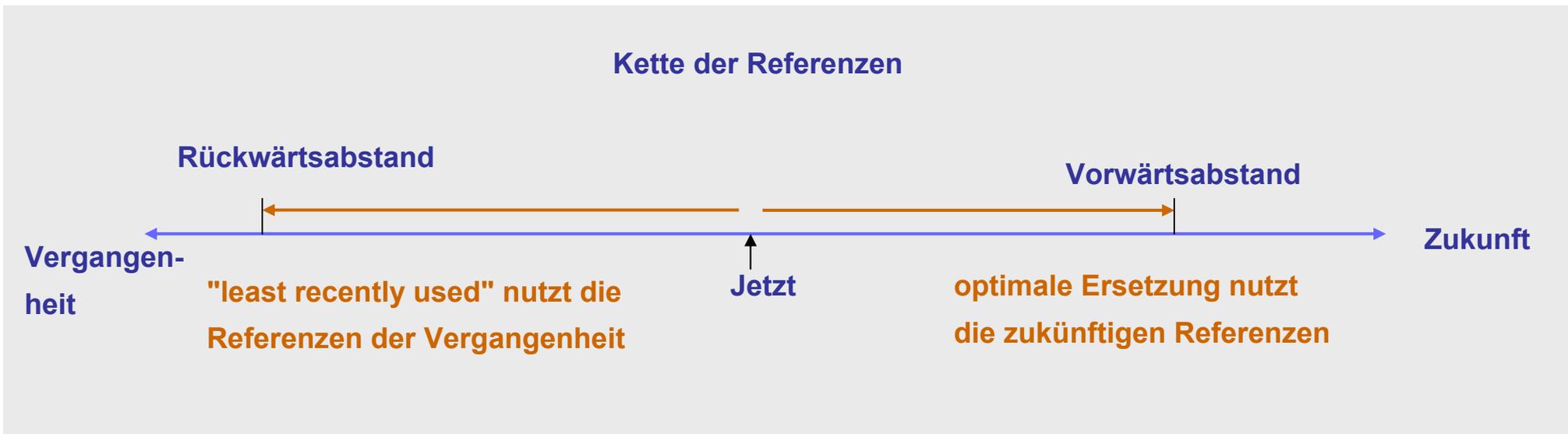
Obwohl mehr Kacheln vorhanden sind, werden mehr Seitenfehler erzeugt!

Grundsätzliches Problem: FIFO berücksichtigt nicht die Nutzung einer Seite. Das gilt auch für 2nd Chance.



Seitenersetzungsstrategien

Least-Recently-Used: Verdränge die Seite, die am längsten nicht referenziert wurde.

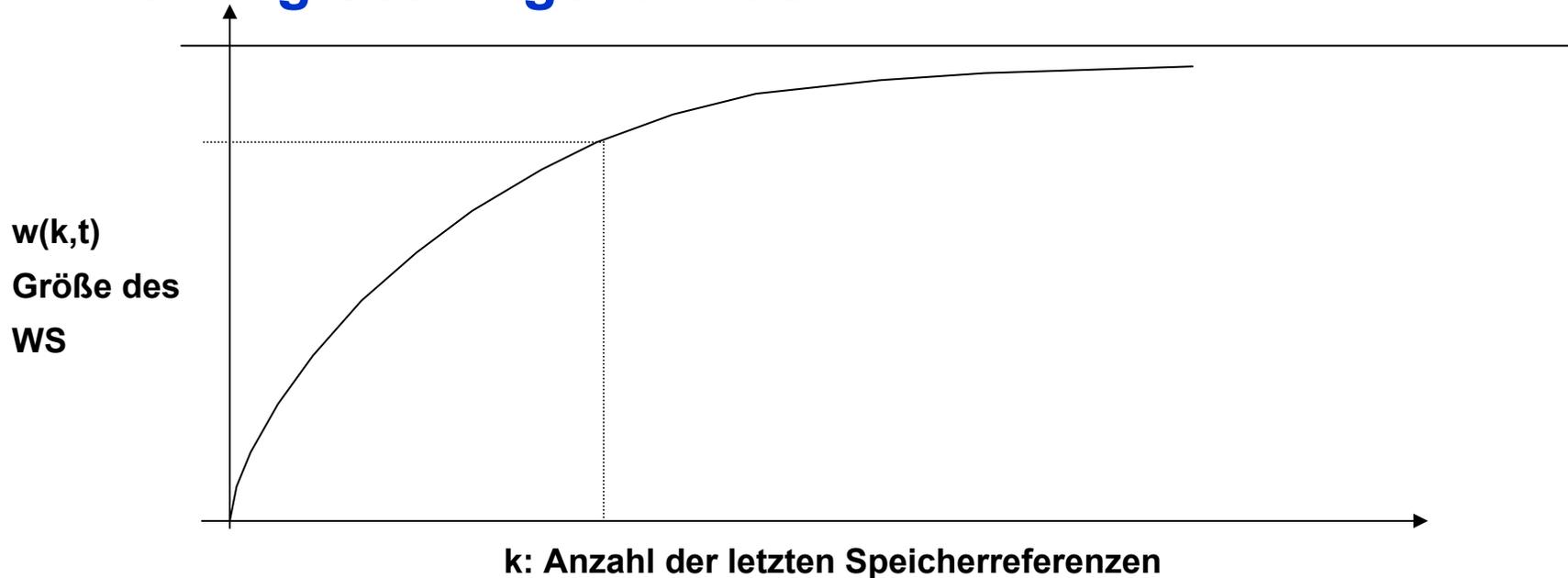


Problem: Least Recently Used ist schwer zu implementieren !



Seitenersetzungsstrategien

Der "Working Set" Algorithmus



Die Menge der Seiten, die von einem Prozess in einem bestimmten Zeitfenster benutzt werden, wird als Working Set (WS) bezeichnet.

Peter Denning: The Working Set Model for Program Behaviour, CACM, May 1968



Der WS Algorithmus

Seitendeskriptor

Verdrängungsalgorithmus:

- durchlaufe alle S.-Deskriptoren
- if $R=1$: set vt to cvt and set $R=0$;
- if $R=0 \wedge (cvt - vt) < t$:
verdränge die Seite;
- if $R=0 \wedge (cvt - vt) > t$:
keine Änderung;
- wenn keine Seite gefunden wird mit:
 $R=0 \wedge (cvt - vt) < t$ then
verdränge älteste Seite;
- wenn alle seiten referenziert wurden
verdränge beliebige Seite.

Seitentabelle	
2083	1
2003	1
1981	1
1212	0
2014	1
2020	1
2032	1
1620	0
.....	

 R-Bit

virtual time: vt

Das Feld enthält die Zeit des letzten Zugriffs auf die Seite.

"virtual time" ist eine Prozess-lokale Repräsentation der Zeit, die mit Prozessbeginn startet.

current virtual time: cvt

2204

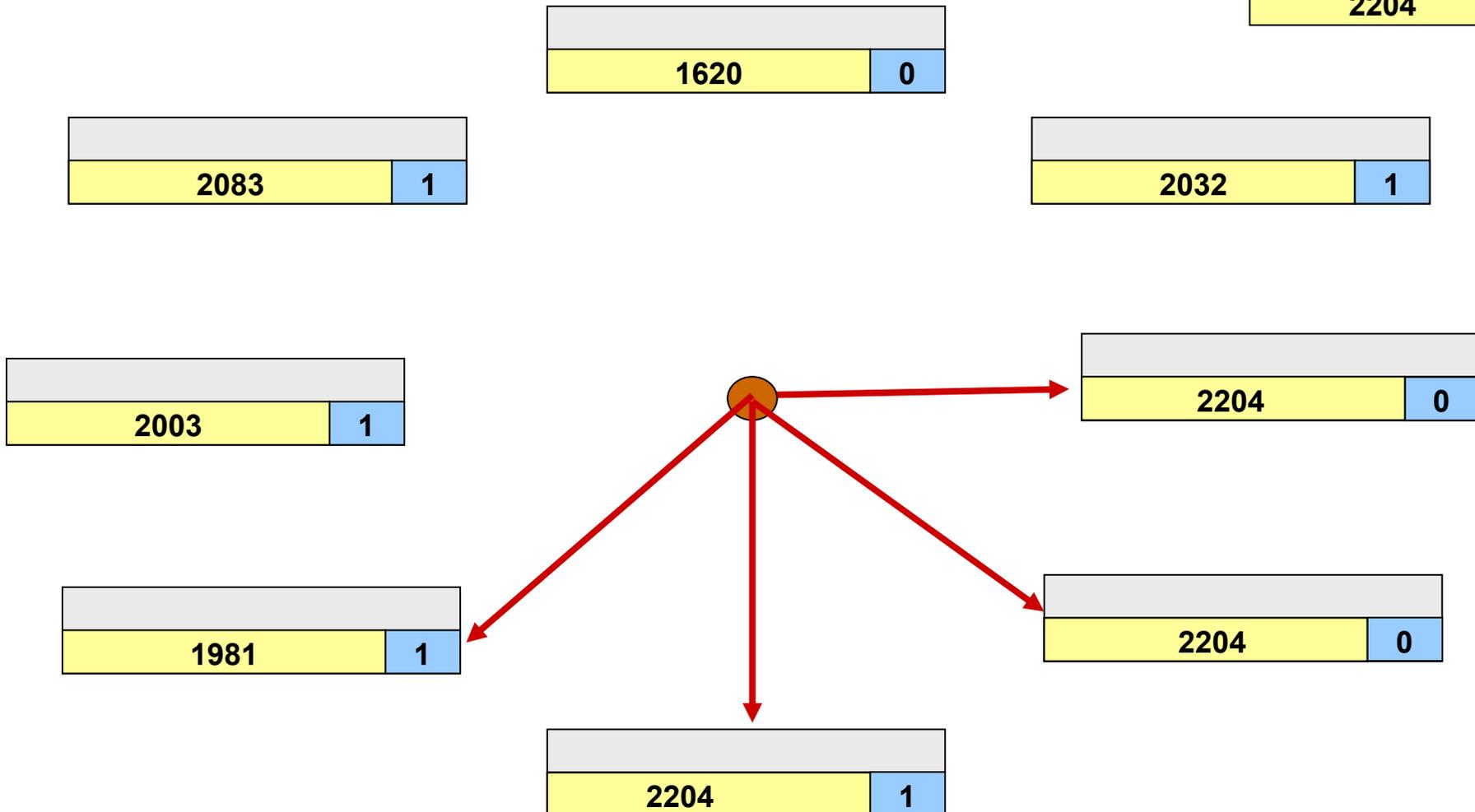


Der WS-Clock Algorithmus

 R-Bit

current virtual time: cvt

 2204



page replacement policies summary

Algorithmus	Eigenschft.	impl.	Kommentar
optimal	😬 😬 😬	😡 😡 😡	nur zum Vergleich, kann nicht realisiert werden
NRU:	😬	😬 😬	einfach und einfach zu realisieren
FIFO	😬	😬 😬	einfach; Problem: wichtige (alte) Seiten werden verdrängt
2nd chance:	😬 😬	😬	substantielle Verbesserung von FIFO
Clock:	😬 😬	😬 😬	intelligente Implementatierung von 2nd Chance
LRU:	😬 😬 😬	😡 😡	exzellent, aber schierig in der Implementierung
WS:	😬 😬 😬	😬	gut, Implementierungsprobleme
WSClock:	😬 😬 😬	😬 😬	gut + effizient

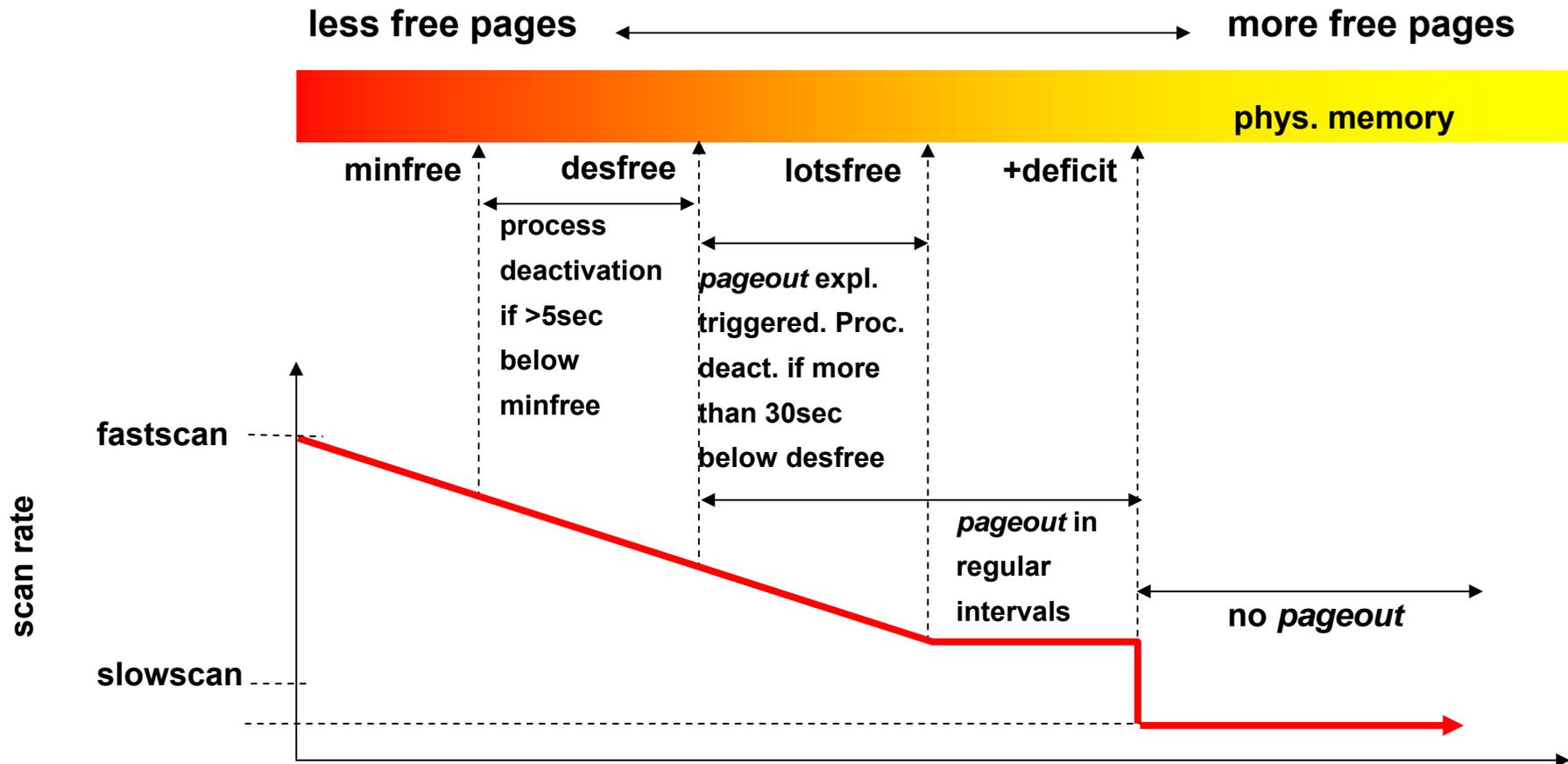


Examples

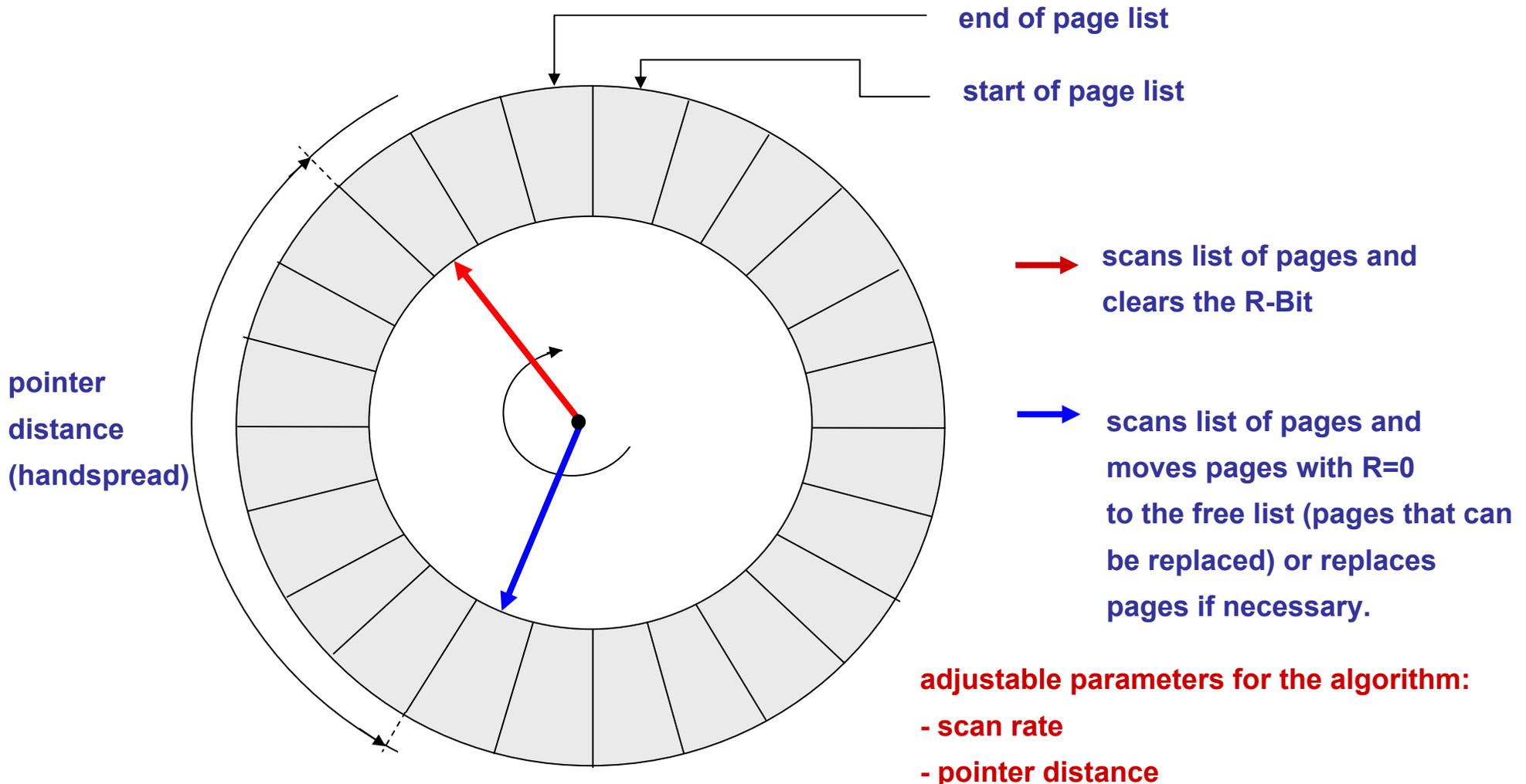
Solaris



Memory management in Unix et al.



releasing pages: clock with two pointers



Examples

Windows 2000



Example: Memory management in Windows 2k

