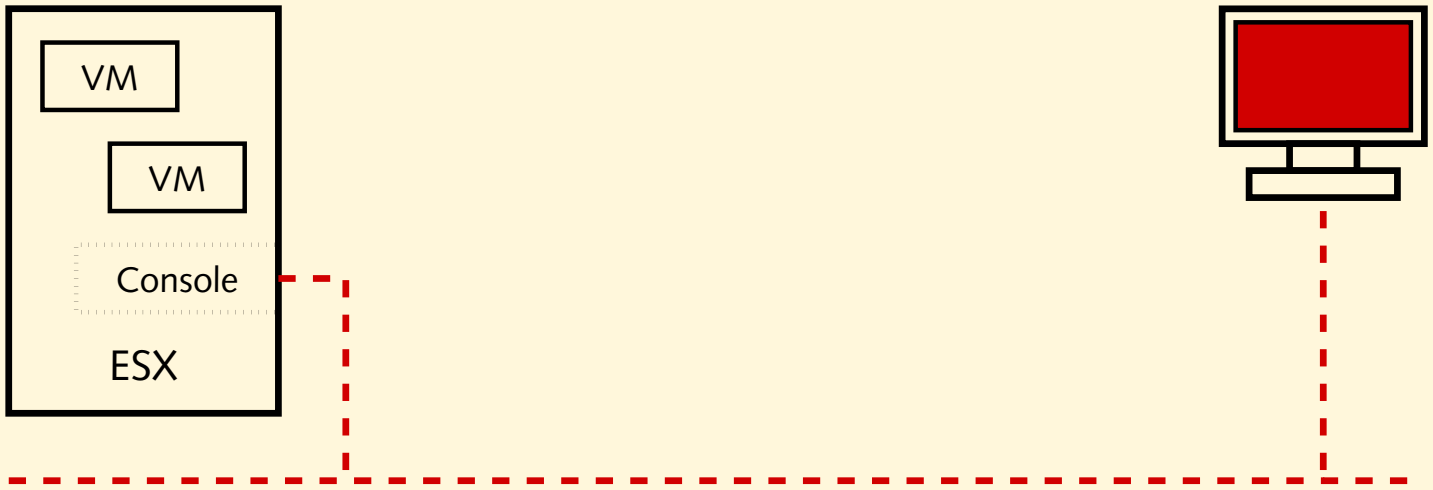
VMware Virtual Infrastructure

Jörn Clausen
joernc@gmail.com

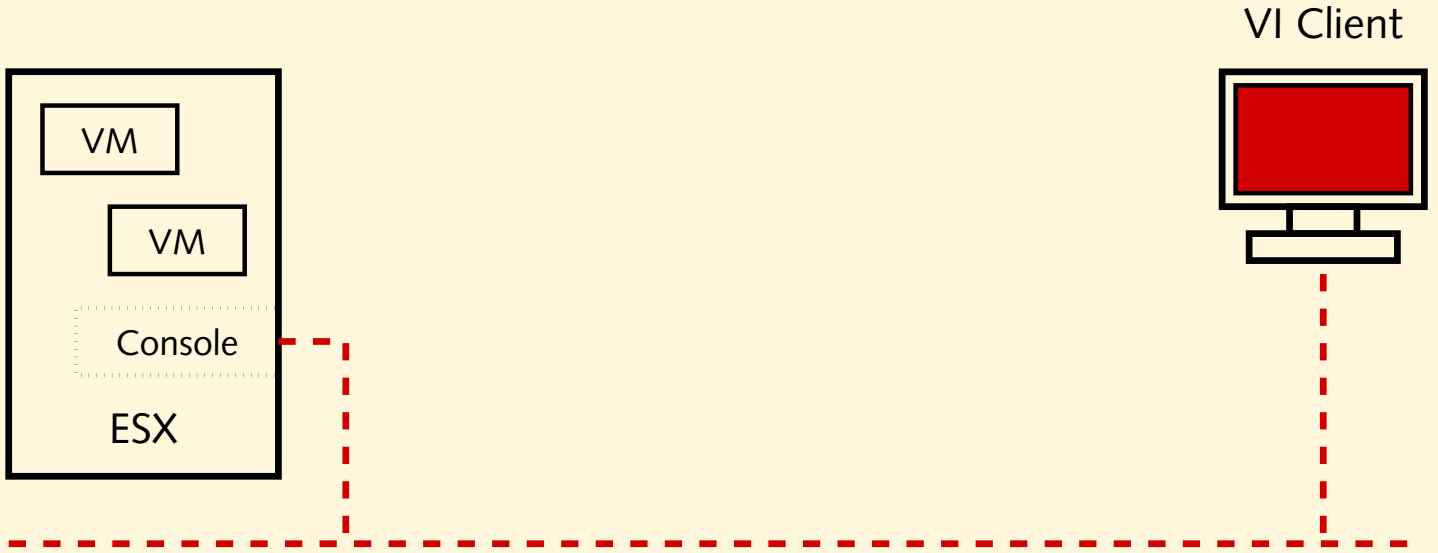
Produktpalette VMware

- verschiedene Hypervisoren:
  - hosted: VMware Server / Workstation / Fusion
  - bare metal: ESX / ESXi
- VMware Infrastructure 3:
  - ESX(i) + Virtual Center + Infrastructure Client
  - Virtual SMP, Update Manager, Consolidated Backup
  - VMotion, Storage VMotion
  - DRS (Dynamic Resource Scheduler)
  - HA (High Availability)
- buzzword für ~~VMI4~~ vSphere 4.0: *Continuous Availability*

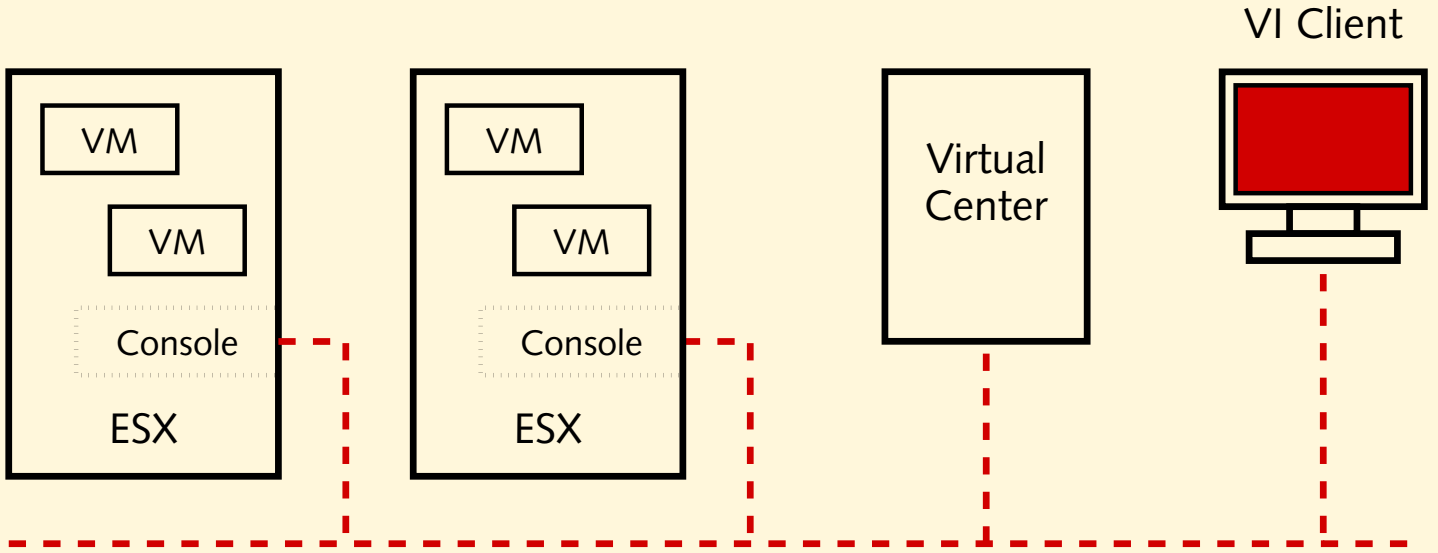
VMware Infrastructure



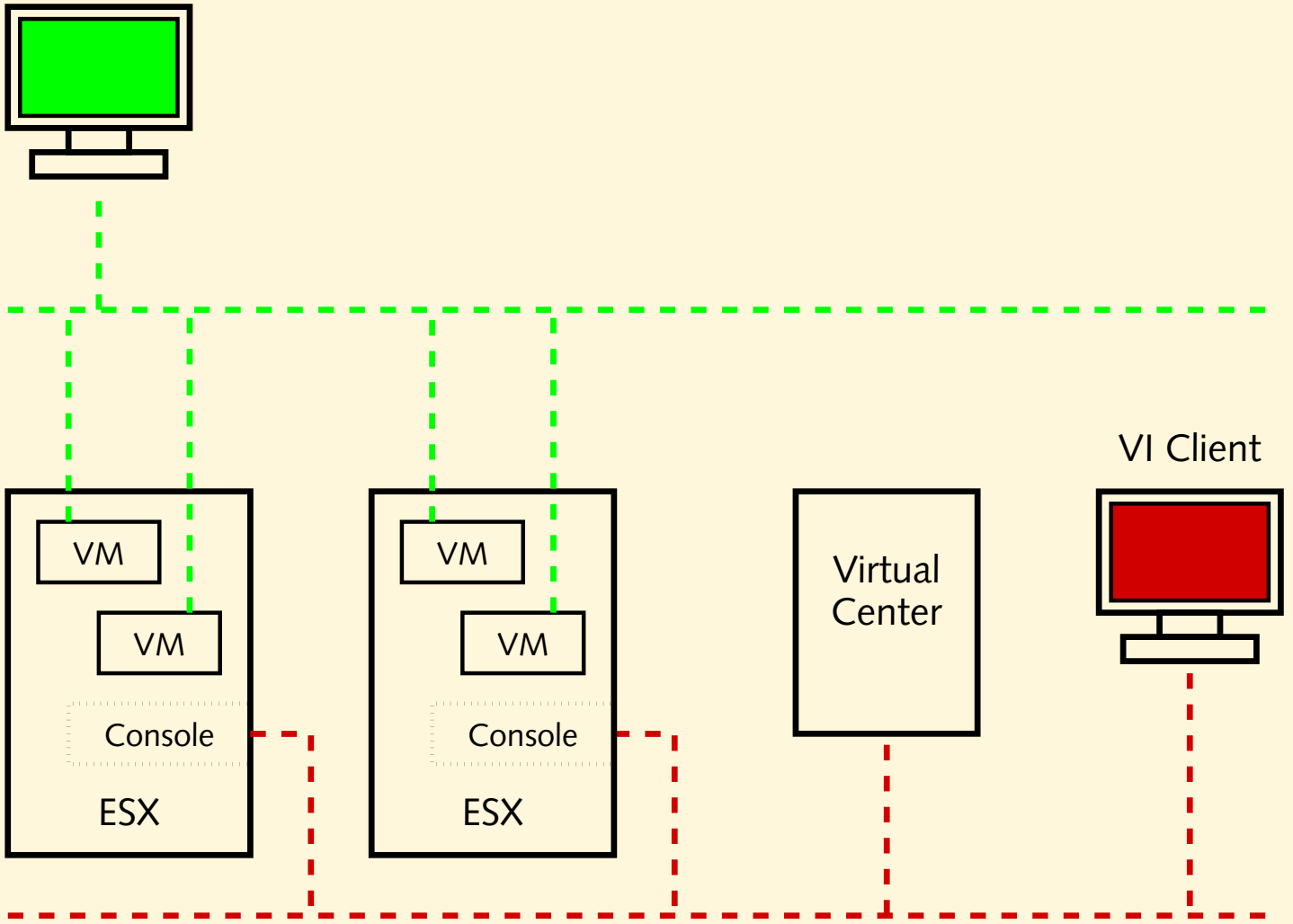
VMware Infrastructure



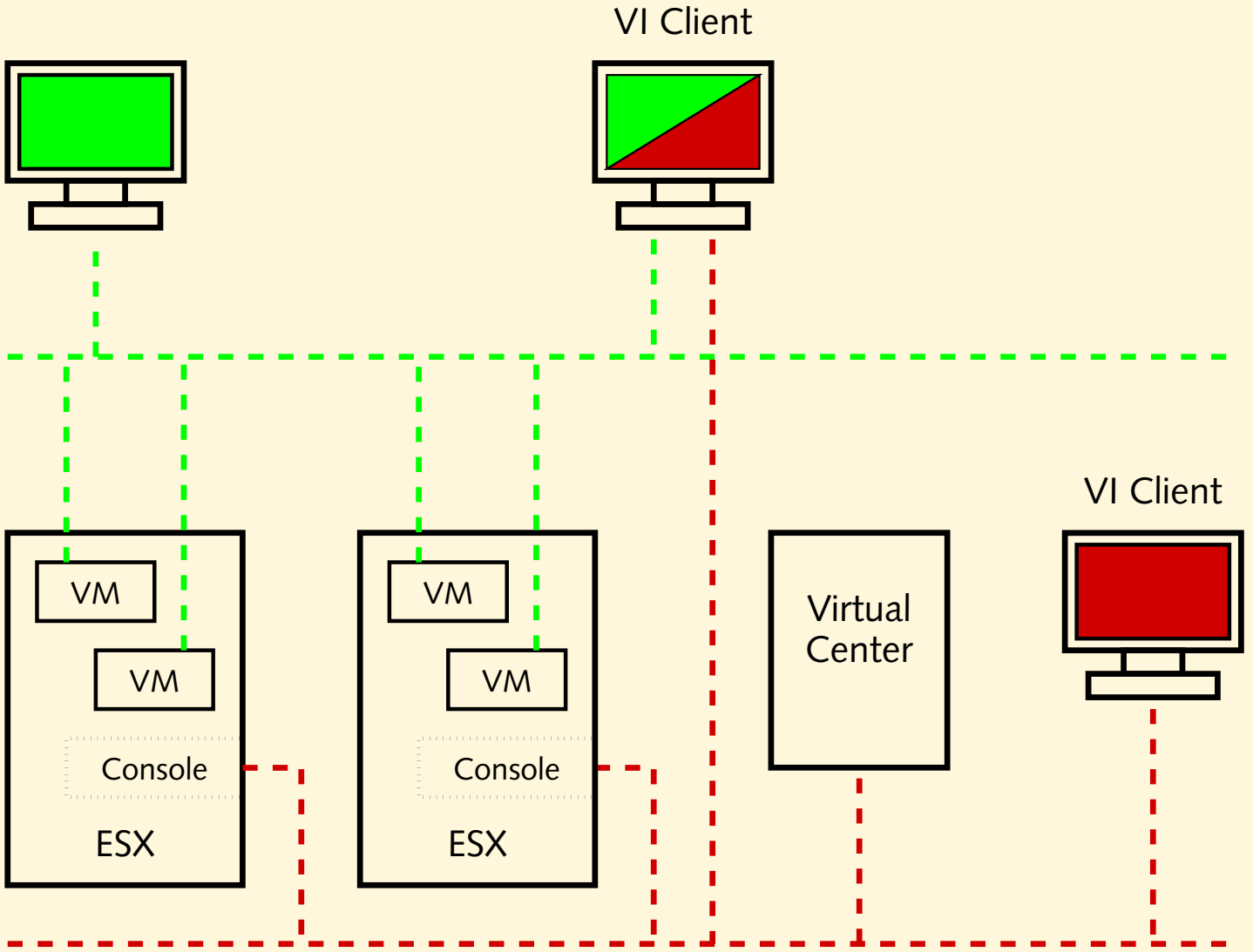
VMware Infrastructure



VMware Infrastructure



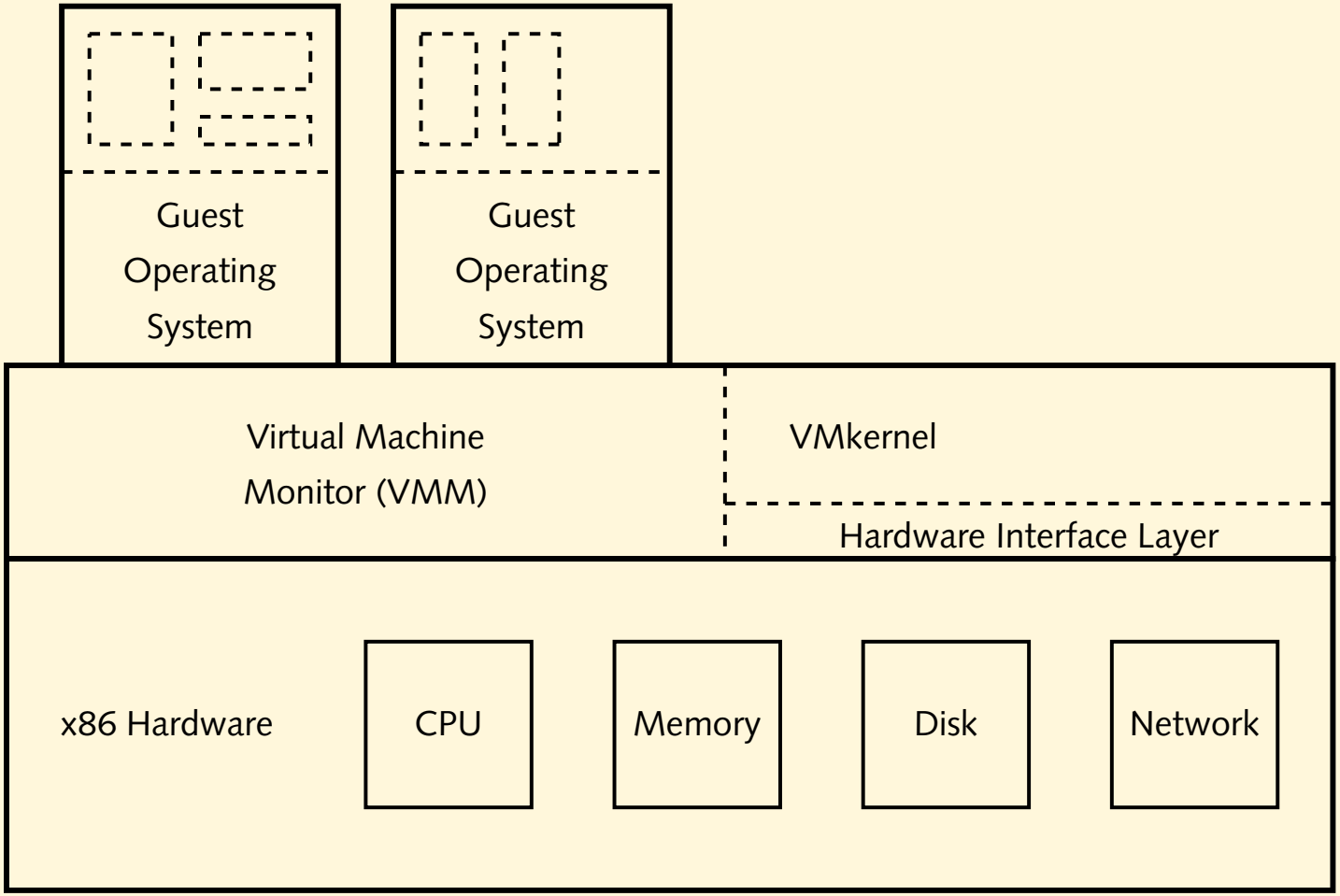
VMware Infrastructure



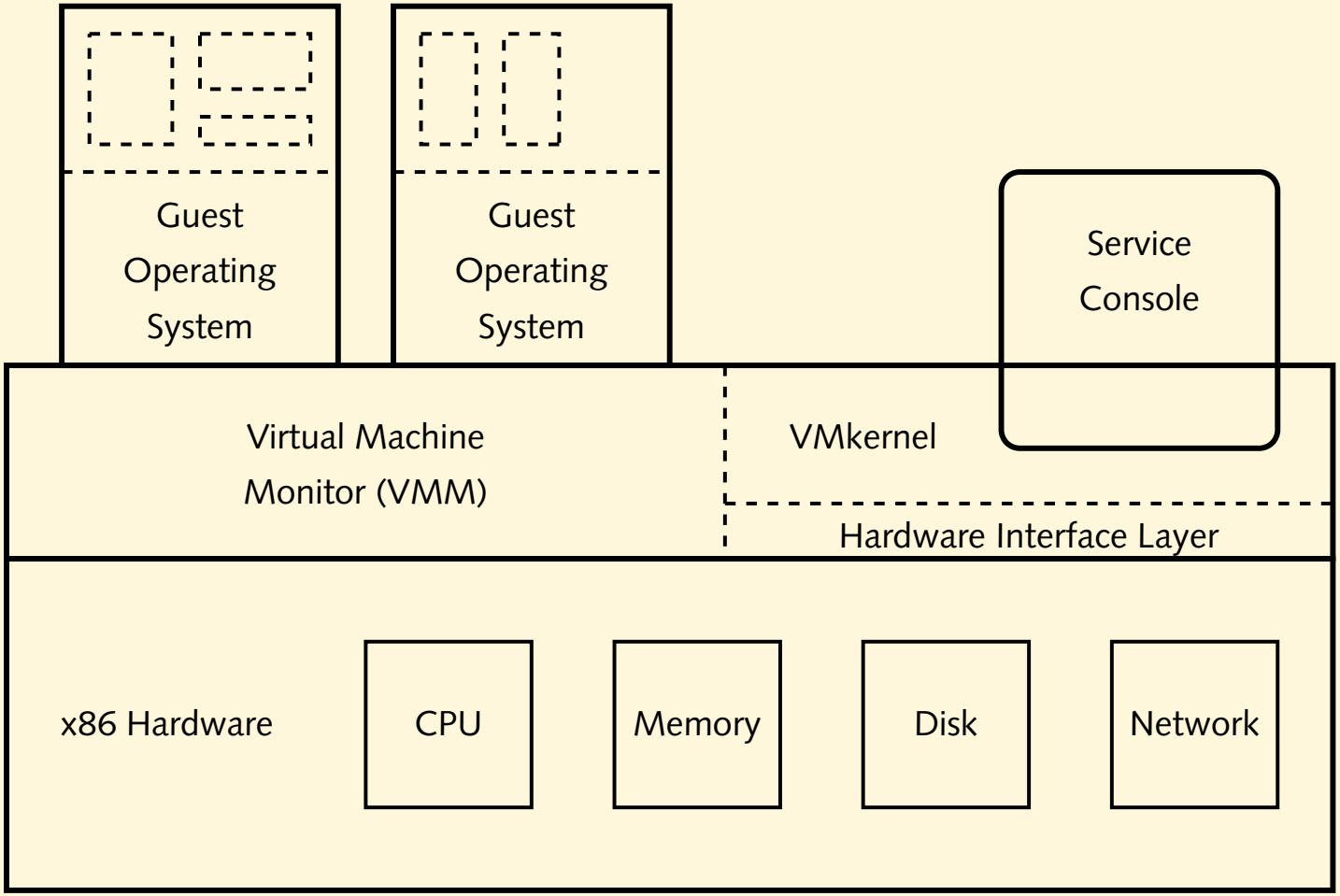
Administration

- Benutzerverwaltung durch
  - ESX-Server (Unix-Accounts)
  - Virtual Center (Windows-Benutzer)
  - Active Directory
- Zuordnung von Rechten zu Rollen
- Zuweisung von Rollen an Benutzer/Gruppen
- Partitionierung des ESX-Clusters in *Resource Pools*
- Zuweisung von CPU- und Speicher-Anteilen

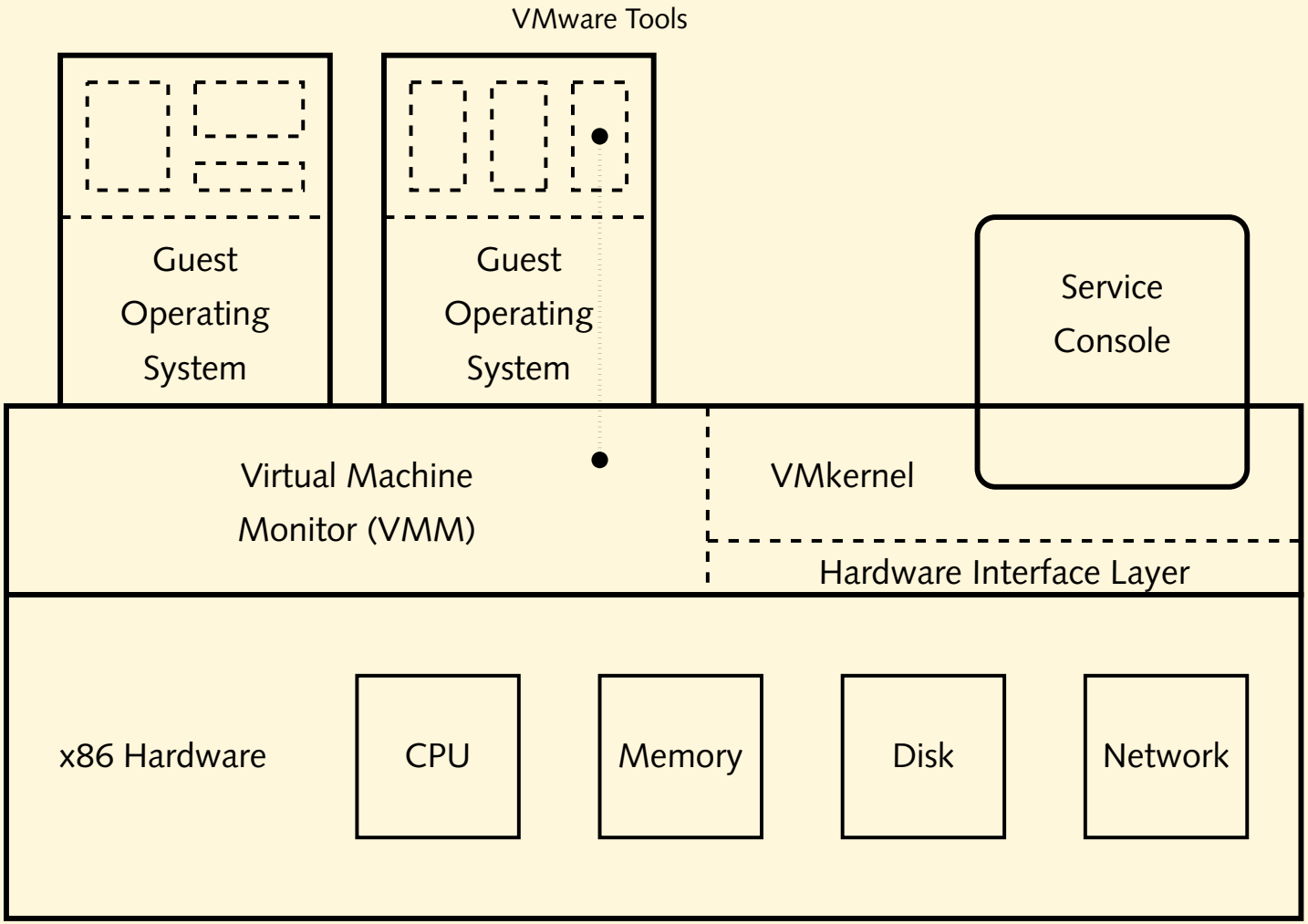
Architektur ESX(i)-Server



Architektur ESX(i)-Server

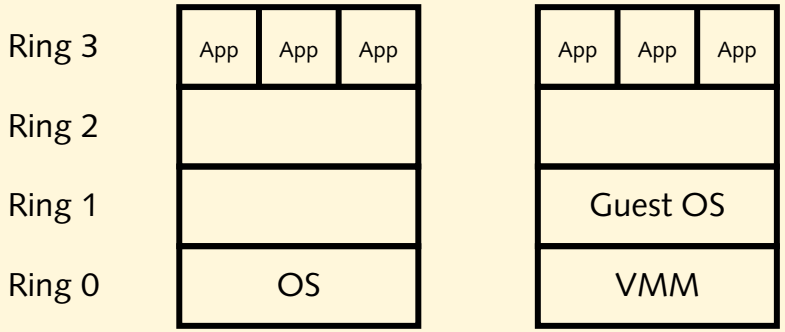


Architektur ESX(i)-Server



Virtualisierung der CPU

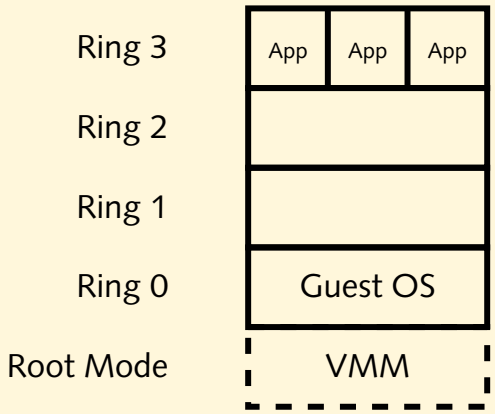
- kernel und userland verteilen sich auf *Ringe*
- Hypervisor verdrängt Gast-Betriebssystem in höheren Ring



- klassischer Virtualisierungsansatz: *trap and emulate*
- *kritische Instruktion* löst trap aus, der von VMM abgefangen wird

Virtualisierung der x86-CPU

- 17 kritische Instruktionen
  - lösen keinen trap aus
  - haben in Ring $\neq$ 0 andere Semantik
- Lösungen:
  - Paravirtualisierung
  - Änderung der x86-Architektur: „Ring -1“ (Intel VT, AMD-V)



  - VMware: *binary translation*

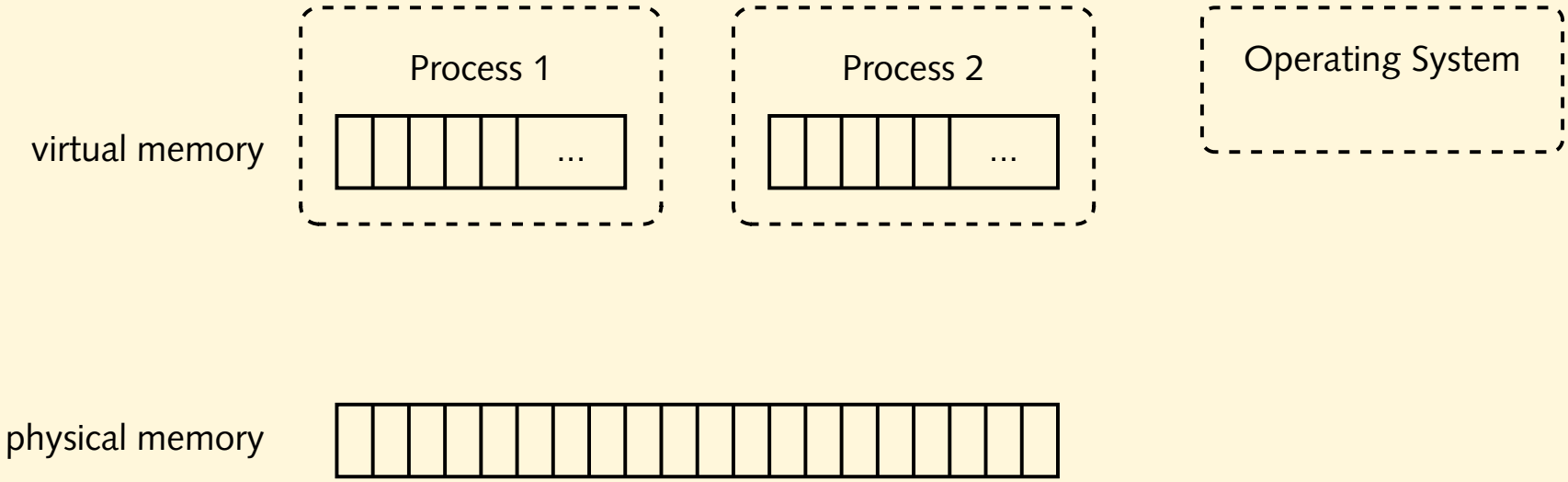
binary translation

- Abfolge von x86-Anweisungen wird in *translation units* aufgeteilt
- Überprüfung auf kritische Instruktionen und Speicherzugriffe
- modifizierter Code wird in *translation cache* abgelegt
- Übersetzung ist *lazy*:
  - bestimmte Teile werden eventuell nie übersetzt (z.B. Fehlerbehandlung)
  - häufig verwendete Teile lassen sich gut cachen
- selbstmodifizierender Code? Vorläufer Embra:
  - Überwachung des Speichers mit Original-Code auf Schreiboperationen
  - komplette Löschung des translation cache bei Änderung
- andere Beispiele für BT: Virtual PC/VirtualBox, JVM, Transmeta Crusoe

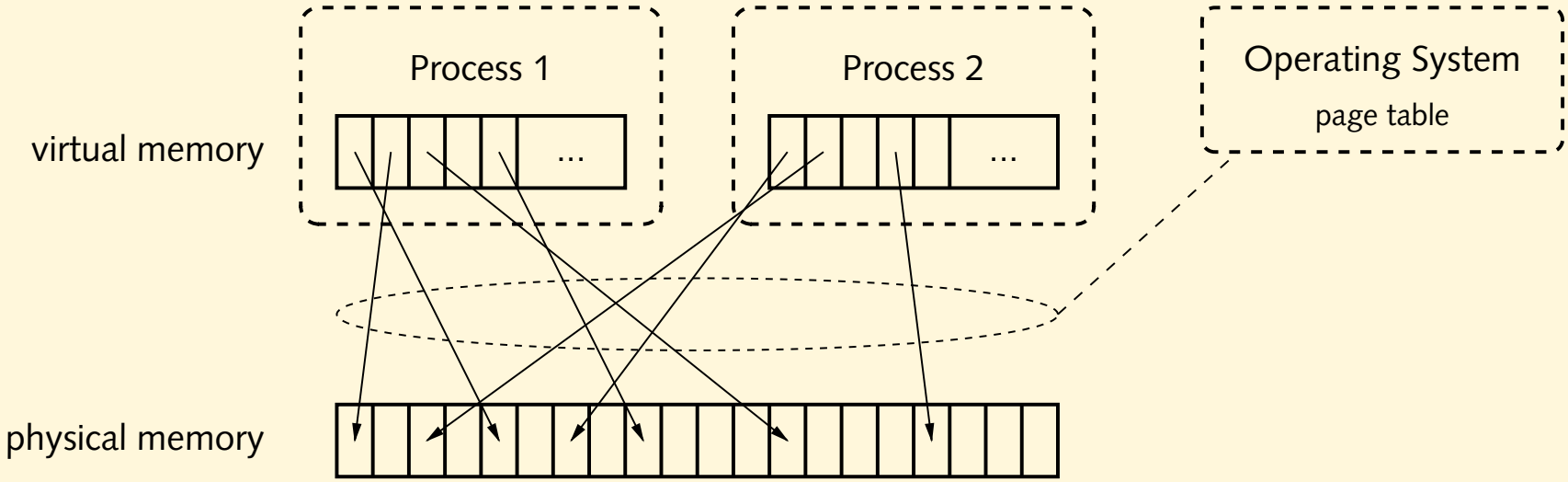
Virtualisierung der Festplatte

- innerhalb der VM: SCSI-Controller (LSI oder BusLogic) mit Festplatten
- Virtual Machine Disk Format (VMDK)
  - snapshots durch *copy-on-write*
  - Dateien innerhalb eines Wirts-Filesystems
- NFS oder VMFS
- Virtual Machine File System (VMFS)
  - lokale Festplatten, Fibre Channel oder iSCSI
  - optimiert für große Dateien
  - gleichzeitiger Zugriff durch mehrere ESX-Server
- Raw Device Mapping
- NAS/SAN ermöglicht einfache Migration einer VM

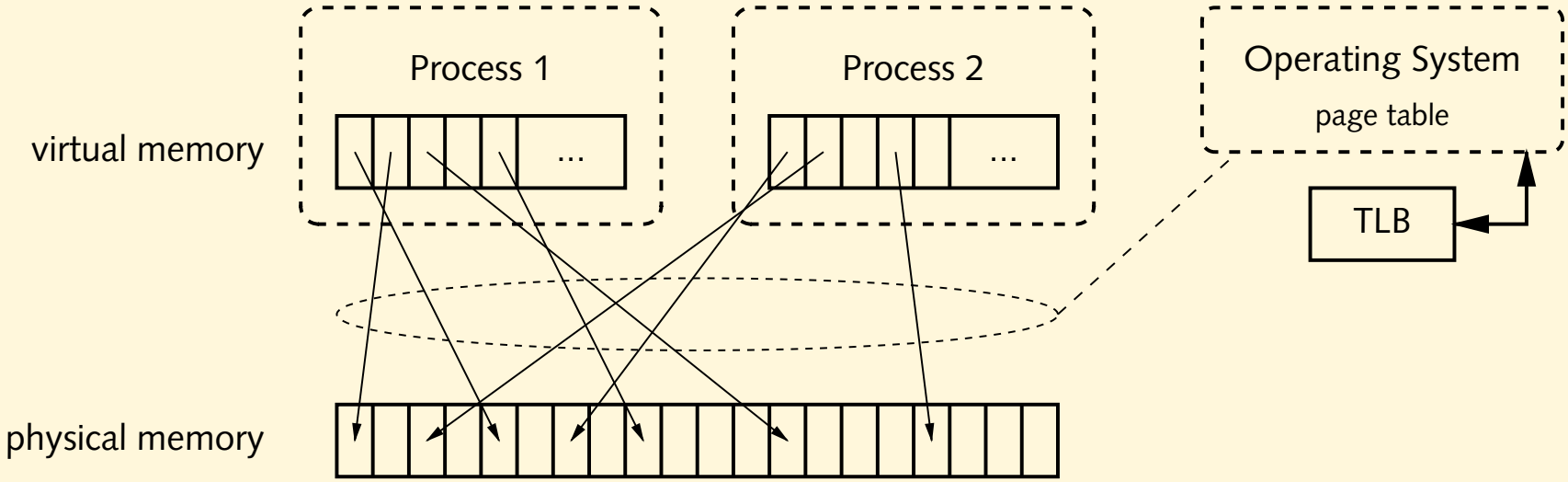
Speicherverwaltung ohne Virtualisierung



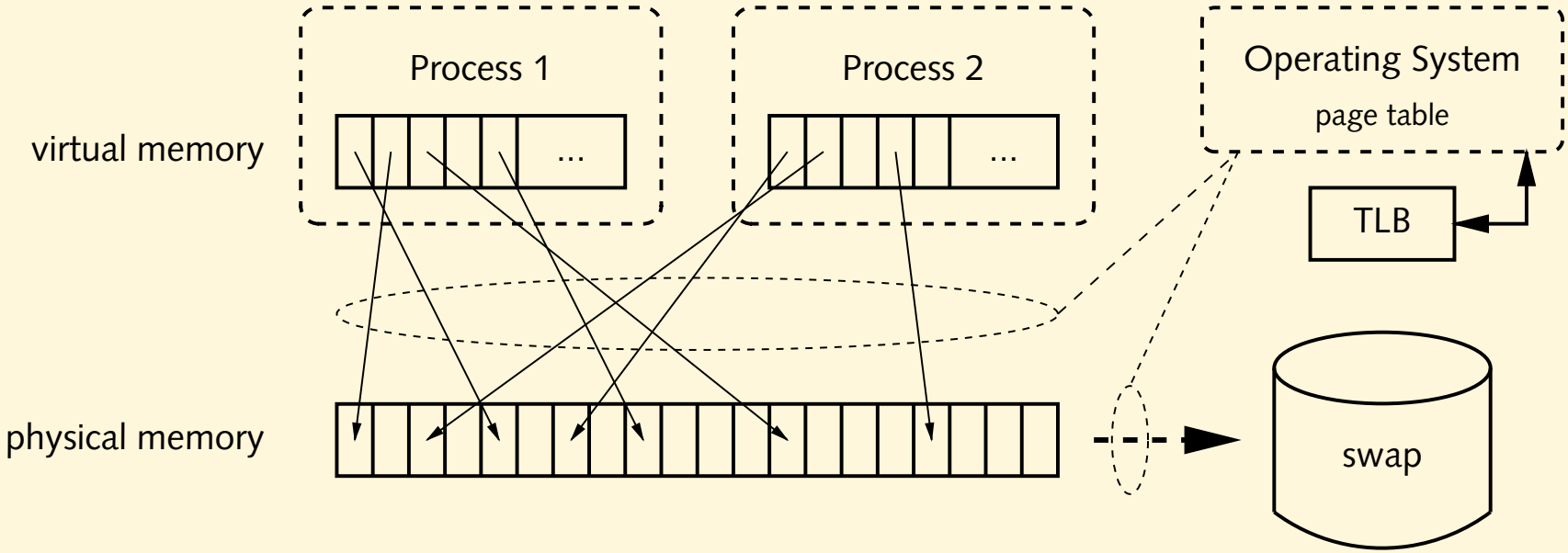
Speicherverwaltung ohne Virtualisierung



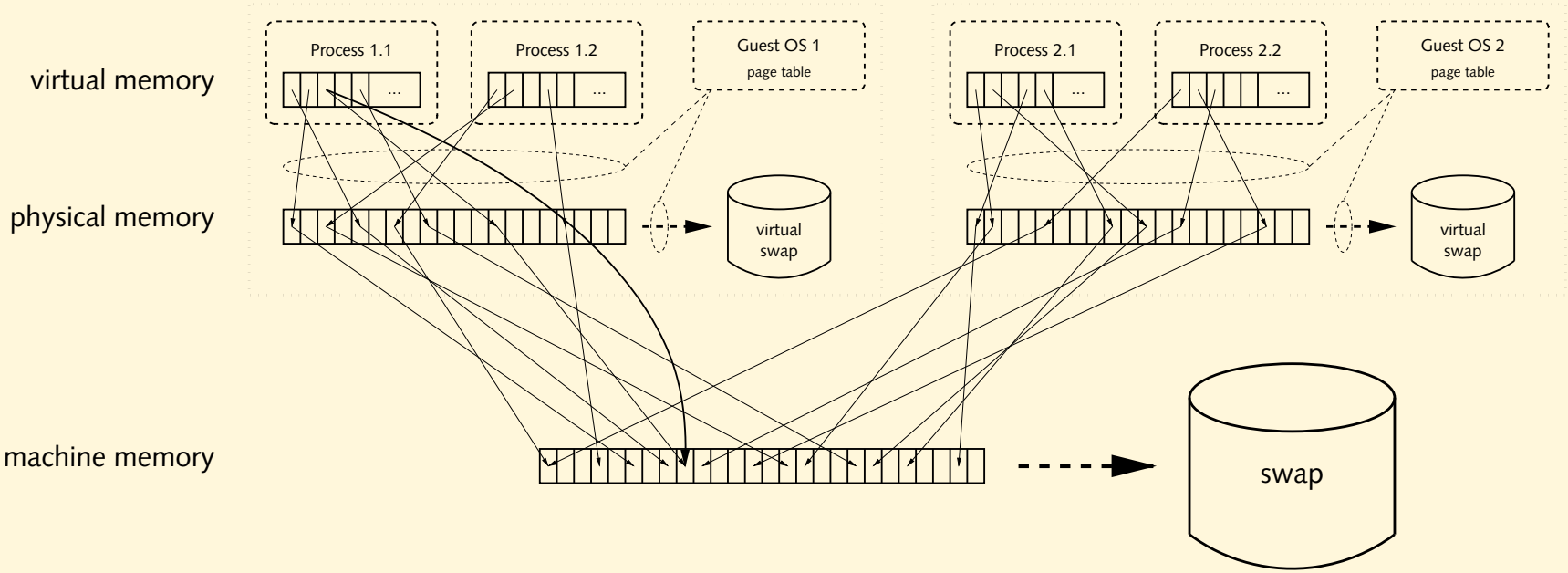
Speicherverwaltung ohne Virtualisierung



Speicherverwaltung ohne Virtualisierung



Speicherverwaltung mit Virtualisierung



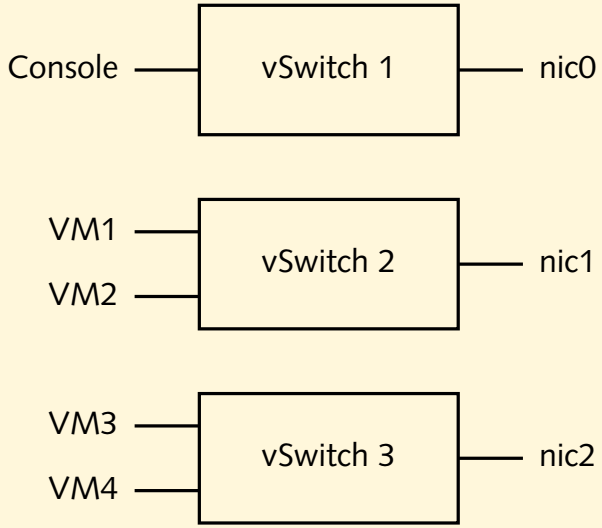
- *machine memory* nicht unbedingt Summe der *physical memories*
- page sharing
- *double paging problem*: swapping des VMM möglichst vermeiden

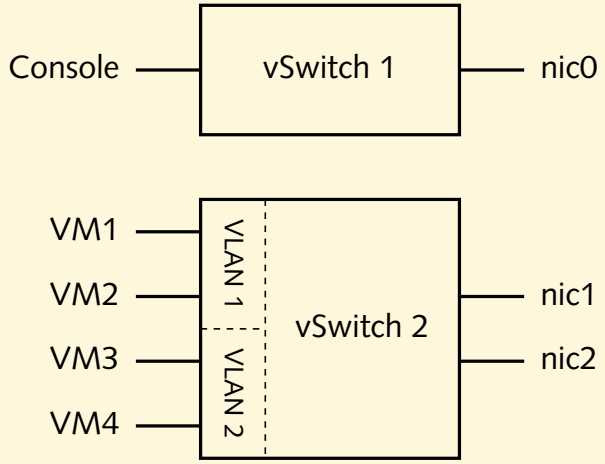
Speicherverwaltung VMware

- derzeit nicht möglich: dynamische Änderung des physikalischen Speichers
- ballooning
  - via VMware Tools wird Speicher innerhalb der VM alloziert
  - Gast-Bestriebssystem beginnt, Speicher auszulagern
  - freigewordener Speicher kann anderen VMs zugewiesen werden
  - Verkleinerung des Ballons, wenn Speicher wieder ausreicht
- idle memory tax
  - tatsächliche Speichernutzung wird durch provozierte page faults ermittelt
  - unbenutzter Speicher wird per ballooning zurückgefordert

Virtualisierung des Netzwerks

- innerhalb der VM: AMD Lance PCNet32, Intel E1000 oder vmxnet
- in der Virtual Infrastructure: vSwitch
- Verbindung von VMs untereinander
- Verbindung von VMs mit physikalischen NICs des ESX-Servers
- NIC teaming
- VLAN tagging

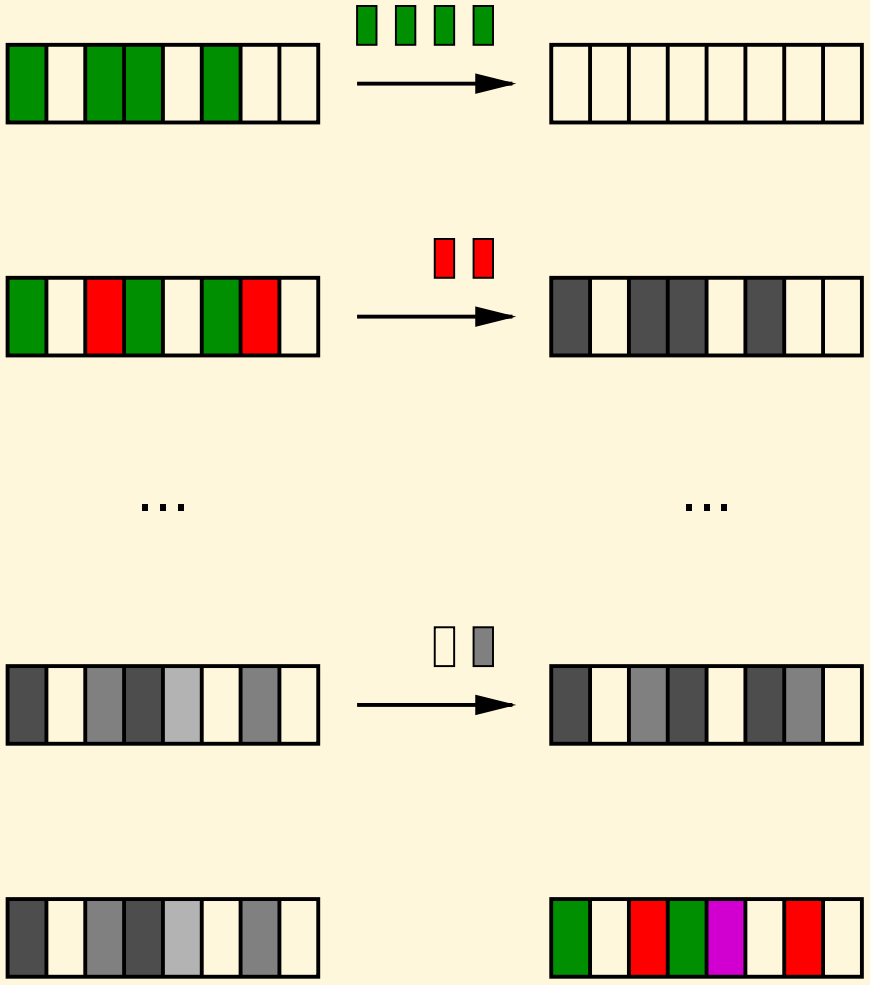




VMotion

- Live-Migration einer VM zwischen zwei ESX-Servern
- Voraussetzungen:
  - identisches Storage-Backend auf beiden ESX-Servern
  - identische Netzwerk-Konfiguration auf beiden ESX-Servern
  - (gleiche Prozessoren in beiden ESX-Servern)
- empfohlen: dediziertes Migrations-Netzwerk
- Ausfallzeit typischerweise unter einer Sekunde
- sollte innerhalb der Toleranz von üblichen Protokollen liegen
- technische Voraussetzungen für ständige Schatten-Kopie einer VM

VMotion, cont.



Transfer aller Seiten

Transfer veränderter Seiten

Wiederholung, bis stabiler Zustand erreicht

Stopp der Quell-VM
Transfer von „working set“ und Meta-Daten

Start der Ziel-VM

Demonstration

Berichte von der Front

- snapshots und Migrations-Techniken
- weitere Leistungseinbußen durch snapshots
- virtuelles SMP
- Delegation von Rechten an „Kunden“

Berichte von der Front

- snapshots und Migrations-Techniken
- weitere Leistungseinbußen durch snapshots
- virtuelles SMP
- Delegation von Rechten an „Kunden“
- **Fazit: Überlege Dir ein Betriebskonzept!!!**

Literatur

- VMware Infrastructure 3 Documentation
  http://www.vmware.com/support/pubs/vi_pages/vi_pubs_35u2.html
- VMware ESX Server 2 – Architecture and Performance Implications
  http://www.vmware.com/pdf/esx2_performance_implications.pdf
- Security Design of the VMware Infrastructure 3 Architecture
  http://www.vmware.com/pdf/vi3_security_architecture_wp.pdf
- Understanding Full Virtualization, Paravirtualization, and Hardware Assist
  http://www.vmware.com/files/pdf/VMware_paravirtualization.pdf
- A Comparison of Software and Hardware Techniques for x86 Virtualization
  http://www.vmware.com/pdf/asplos235_adams.pdf
- Analysis of the Intel Pentium's Ability to Support a Secure Virtual Machine Monitor
  http://www.usenix.org/events/sec2000/robin.html
- Embra: Fast and Flexible Machine Simulation
  http://www.cs.utexas.edu/users/witchel/pubs/SIGMetrics96-embra.pdf
- Virtualizing I/O Devices on VMware Workstations's Hosted Virtual Machine Monitor
  http://www.usenix.org/publications/library/proceedings/usenix01/sugerman.html

Literatur, cont.

- Memory Resource Management in VMware ESX Server
  http://www.usenix.org/publications/library/proceedings/osdi02/tech/waldspurger.html
- The Role of Memory in VMware ESX Server 3
  http://www.vmware.com/pdf/esx3_memory.pdf
- VMware Virtual Networking Concepts
  http://www.vmware.com/files/pdf/virtual_networking_concepts.pdf
- Fast Transparent Migration for Virtual Machines
  http://www.usenix.org/events/usenix05/tech/general/nelson.html
- Live Migration of Virtual Machines
  http://www.usenix.org/events/nsdi05/tech/clark.html
- Xen und VMware ESX: Sicherheitsprobleme in x86-Virtualisierungsumgebungen
  Best Practice #24: Absicherung von VMware ESX Servern
  http://www.gai-netconsult.de/de/download/security/secjournal/index.html