7 Heutige asymmetrische Verfahren

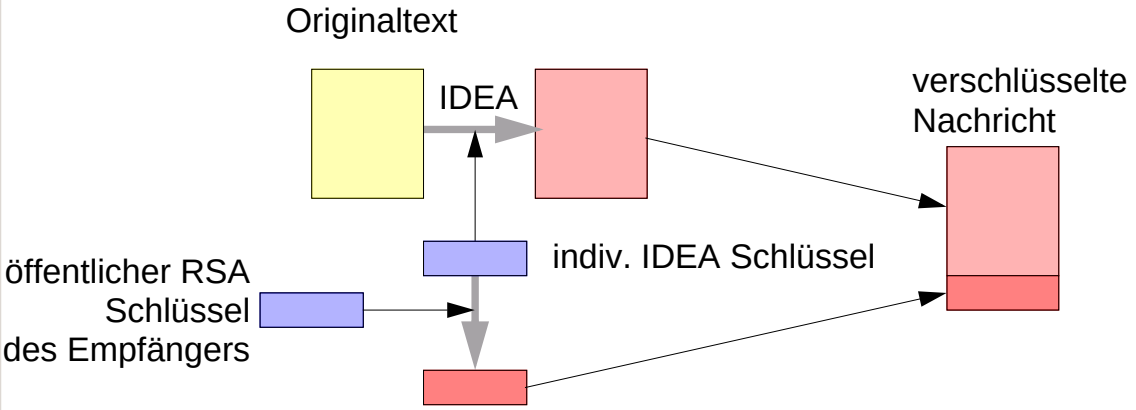
- RSA (Rivest, Shamir und Adleman)
  - Öffentlicher Schlüssel (zum Verschlüsseln) besteht aus *(e, N)*
  - Ein Block *M* wird verschlüsselt durch: $C = E((e, N), M) = M^e \bmod N$
  - *C* wird entschlüsselt durch: $M = D((d, N), C) = C^d \bmod N$
  - Wahl der Schlüssel:
    - Es muß gelten $\forall M: (M^e)^d = M \bmod N$
    - Aus Kenntnis von *e* und *N* darf *d* nur mit hohem Aufwand ermittelbar sein
  - Lösung:
    - $N = pq$ mit *p* und *q* zwei hinreichend große Primzahlen
    - zufällige Wahl von *d*, teilerfremd zu $(p-1)(q-1)$
    - Berechnung von *e* aus der Bedingung: $ed = 1 \bmod ((p-1)(q-1))$
  - Es ist aufwendig, die Primfaktoren von *N* zu berechnen (mit diesen wäre es möglich *d* zu ermitteln)

7 Heutige asymmetrische Verfahren (2)

- Vorteil asymmetrischer Verfahren (*Public key*-Verfahren)
  - nur ein Schlüsselpaar pro Teilnehmer nötig
    (sonst ein Schlüsselpaar pro Kommunikationskanal!)
  - Schlüsselverwaltung erheblich vereinfacht
    - jeder Teilnehmer erzeugt sein Schlüsselpaar und
    - veröffentlicht seinen öffentlichen Schlüssel
  - Authentisierung durch digitale Unterschriften möglich
  - gilt als sicher bei hinreichend großer Schlüssellänge (1024 Bit)
- Nachteil
  - relativ langsam berechenbar
  - gemischter Betrieb von asymmetrischen und symmetrischen Verfahren zur Geschwindigkeitssteigerung

7 Heutige asymmetrische Verfahren (3)

- Beispiel: PGP Verschlüsselung



  - ◆ Daten werden mit einem individuellen Schlüssel IDEA-verschlüsselt
  - ◆ IDEA-Schlüssel wird RSA-verschlüsselt der Nachricht angehängt

- ★ Nachricht an mehrere Adressaten verschickbar
  - ◆ lediglich der IDEA-Schlüssel muß in mehreren Varianten verschickt werden (je eine Version verschlüsselt mit dem öffentl. Schlüssel des Empfängers)

8 Digitale Unterschriften

- Authentisierung des Absenders
  - ◆ Bilden eines Hash-Wertes über die zu übermittelnde Nachricht
    - Hash-Wert ist ein Codewort fester Länge
    - es ist unmöglich oder nur mit hohem Aufwand möglich, für einen gegebenen Hash-Wert eine zugehörige Nachricht zu finden
  - ◆ Verschlüsseln des Hash-Wertes mit dem geheimen Schlüssel des Absenders (digitale Unterschrift, digitale Signatur)
  - ◆ Anhängen des verschlüsselten Hash-Wertes an die Nachricht
  - ◆ Empfänger kann den Hash-Wert mit dem öffentlichen Schlüssel des Absenders dechiffrieren und mit einem selbst berechneten Hash-Wert der Nachricht vergleichen
  - ◆ stimmen beide Werte überein muß die Nachricht vom Absender stammen, denn nur der besitzt den geheimen Schlüssel

8 Digitale Unterschriften (2)

- Kombination mit Verschlüsselung
  - erst signieren
  - dann mit dem öffentlichen Schlüssel des Adressaten verschlüsseln
- Reihenfolge wichtig
  - Man signiere nichts, was man nicht entschlüsseln kann
- Heute gängiges Hash-Verfahren
  - MD5
  - 128 Bit langer Hash-Wert

8 Digitale Unterschriften (3)

- Woher weiß ich, daß ein öffentlicher Schlüssel authentisch ist?
  - Ich bekomme den Schlüssel vom Eigentümer (persönlich, telefonisch).
    - Hash-Wert auf öffentlichen Schlüsseln, die leichter zu überprüfen sind (Finger prints)
  - Ich vertraue jemandem (Bürge), der zusichert, daß der Schlüssel authentisch ist.
    - Schlüssel werden von dem Bürgen signiert.
    - Bürge kann auch eine ausgezeichnete Zertifizierungsstelle sein.
    - Netzwerk von Zusicherungen auf öffentliche Schlüssel *(Web of Trust)*
  - Möglichst weite Verbreitung von öffentlichen Schlüsseln ereichen (z.B. PGP: Webserver als Schlüsselserver)

8 Digitale Unterschriften (4)

- Mögliche Probleme von Public key-Verfahren
  - Geheimhaltung des geheimen Schlüssels
    (Time sharing-System, Backup; Schlüsselpaßwort / Pass phrase)
  - Vertrauen in die Programme (z.B. PGP)
  - Ausspähung während des Ver- und Entschlüsselungsvorgangs

I.7 Authentisierung im Netzwerk

- Viele Klienten, die viele Dienste in Anspruch nehmen wollen
  - Dienste *(Server)* wollen wissen welcher Benutzer *(Principal)*, den Dienst in Anspruch nehmen will (z.B. zum Accounting, Zugriffsschutz, etc.)
  - Im lokalen System reicht die (durch das Betriebssystem) geschützte Benutzerkennung (z.B. UNIX UID) als Ausweis
  - Im Netzwerk können Pakete abgefangen, verfälscht und gefälscht werden (einfache Übertragung einer Benutzerkennung nicht ausreichend sicher)
- Public key-Verfahren
  - Authentisierung durch digitale Unterschrift (mit geheimen Schlüssel des Senders) und Verschlüsseln (mit öffentlichem Schlüssel des Empfängers)
  - Nachteile
    - jeder Dienst benötigt sicheren Zugang zu allen öffentlichen Schlüsseln
    - Verschlüsseln und Signieren mit RSA ist sehr teuer

I.7 Authentisierung im Netzwerk (2)

★ Einsatz von Authentisierungsdiensten

- ◆ zentraler Server, der alle Benutzer kennt
- ◆ Authentisierungsdienst garantiert einem Netzwerkdienst, daß ein Benutzer auch der ist, der er vorgibt zu sein

■ Benutzerausweis

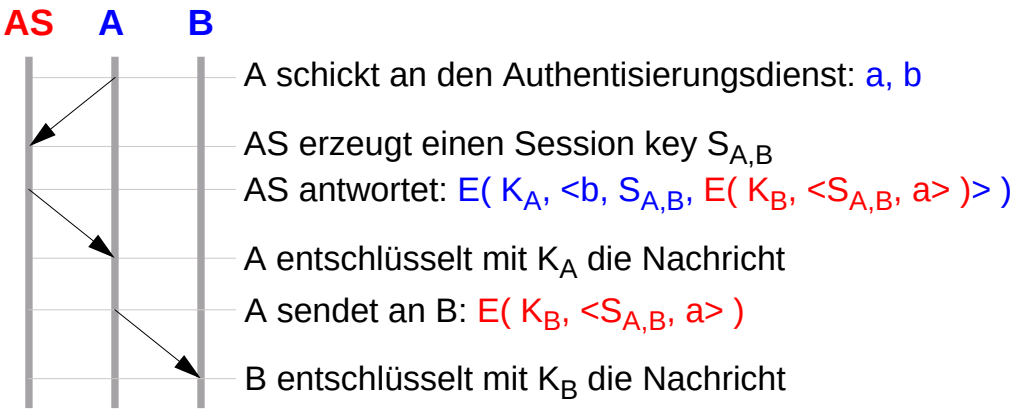
- ◆ der Authentisierungsdienst erkennt den Benutzer anhand eines geheimen Schlüssels oder Paßworts
- ◆ der Schlüssel ist nur dem Authentisierungsdienst und dem Benutzer bekannt

■ Vorgang

- ◆ Benutzer *(Principal)* will mit einem Programm *(Client)* einen Dienst *(Server)* in Anspruch nehmen
- ◆ durch geeignetes Protokoll erhalten Client und Server jeweils einen nur ihnen bekannten Schlüssel, mit dem sie ihre Kommunikation verschlüsseln können *(Session key)*

1 Einfacher Authentisierungsdienst

■ A will den Dienst B in Anspruch nehmen (Nach Needham-Schröder):



- ◆ K_X ist der geheime Schlüssel, den nur Authentisierungsdienst und X kennen
- ◆ nach dem Protokollablauf kennen sowohl A und B den Session key
- ◆ A weiß, daß nur B den Session key kennt
- ◆ B weiß, daß nur A den Session key kennt

1 Einfacher Authentisierungsdienst (2)

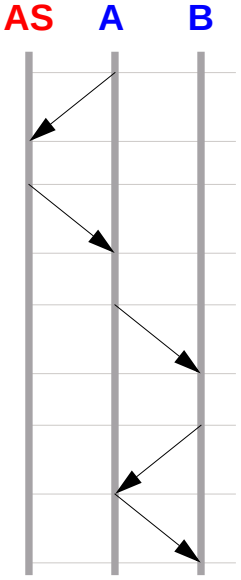
▲ Problem

- ◆ letzte Nachricht von A an B könnte aufgefangen und später erneut ins Netz gegeben werden *(Replay attack)*
- ◆ Folge: Kommunikation zwischen A und B kann gestört werden

★ Korrektur durch zusätzliches Versenden einer Verbindungsbestätigung durch B und A

2 Authentisierungsdienst mit Bestätigung

■ Bestätigung enthält Einmalinformation *(Nonce)*



A schickt an den Authentisierungsdienst: a, b, i_a

AS erzeugt einen Session key $S_{A,B}$

AS antwortet: $E(K_A, <i_a, b, S_{A,B}, E(K_B, <S_{A,B}, a>)>)$

A entschlüsselt mit K_A die Nachricht

A sendet an B: $E(K_B, <S_{A,B}, a>)$

B entschlüsselt mit K_B die Nachricht

B schickt an A: $E(S_{A,B}, i_b)$

A antwortet an B: $E(S_{A,B}, i_b+1)$

- ◆ ein Wiedereinspielen der Nachricht $E(K_B, <S_{A,B}, a>)$ oder $E(S_{A,B}, i_b+1)$ wird erkannt und kann ignoriert werden

2 Authentisierungsdienst mit Bestätigung (2)

▲ Problem

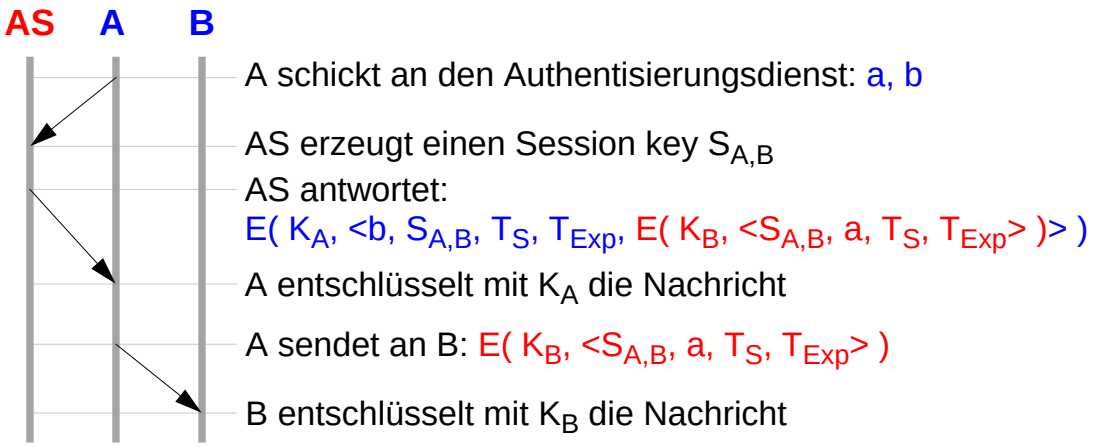
- ◆ Aufzeichnen von $E(K_B, <S_{A,B}, a>)$ und
- ◆ Brechen von $S_{A,B}$ erlaubt das Aufbauen einer Verbindung.
- ◆ ein Dritter kann dann die erste Bestätigung abfangen und die zweite Bestätigung verschicken

★ Lösung

- ◆ Einführung von Zeitstempeln *(Time stamp)* und Angaben zur Lebensdauer *(Expiration time)*

3 Authentisierungsdienst mit Zeitstempeln

■ Authentisierungsdienst versieht seine Nachrichten mit Zeitstempeln



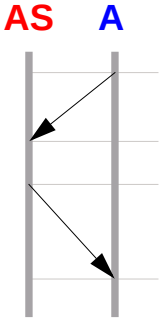
- ◆ T_S = Zeitstempel der Nachrichtenerzeugung
- ◆ T_{Exp} = maximale Lebensdauer der Nachricht
- ◆ aufgezeichnete Nachricht kann nach kurzer Zeit (z.B. 5min) nicht nocheinmal zum Aufbau einer Verbindung verwendet werden

4 Beispiel: Kerberos

- Kerberos V5
  - Softwaresystem implementiert Weiterentwicklung des Needham-Schröder-Protokolls
  - entwickelt am MIT seit 1986
- Ziel
  - Authentisierung und Erzeugung eines gemeinsamen Schlüssels durch den vertrauenswürdigen Kerberos-Server
- Idee
  - Trennung von Authentisierungsdienst und Schlüsselerzeugung
  - reduziert die nötige Übertragung einer Identifikation oder eines Paßworts zum Kerberos-Server

4 Beispiel: Kerberos (2)

- Benutzer holt sich zunächst ein Ticket vom Authentisierungsdienst



A schickt an den Authentisierungsdienst: a, tgs, T_{Exp}, i_a

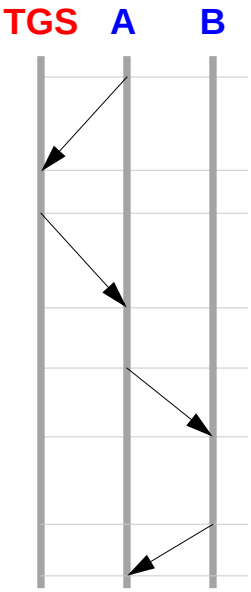
AS erzeugt ein Ticket für den Ticket Server tgs
AS antwortet:
$E(K_A, <S_{A,TGS}, tgs, T_{Exp}, i_a, E(K_{TGS}, <S_{A,TGS}, a, T_{Exp}>)>)$

A entschlüsselt mit K_A die Nachricht

  - das Ticket besteht aus $<S_{A,TGS}, a, T_{Exp}>$
  - es enthält einen Session key für die Kommunikation mit einem Ticket granting server, der dann die Verbindung zu einem Netzwerkdienst bereitstellen kann

4 Beispiel: Kerberos (3)

- Authentisierungsdienst versieht seine Nachrichten mit Zeitstempeln



A schickt an den Ticket granting server:
$E(S_{A,TGS}, T_s), E(K_{TGS}, <S_{A,TGS}, a, T_{Exp}>), b, T_{Exp}, i_a'$

TGS erzeugt einen Session key $S_{A,B}$

TGS antwortet:
$E(S_{A,TGS}, <S_{A,B}, b, T_{Exp}, i_a'>), E(K_B, <S_{A,B}, a, T_{Exp}>)$

A entschlüsselt mit $S_{A,TGS}$ die Nachricht

A sendet an B: $E(S_{A,B}, <T_s, C, S_{sub}>), E(K_B, <S_{A,B}, a, T_{Exp}>)$

B entschlüsselt mit K_B den zweiten Teil der Nachricht (Ticket) und mit $S_{A,B}$ den ersten Teil

B antwortet (optional): $E(S_{A,B}, T_s)$

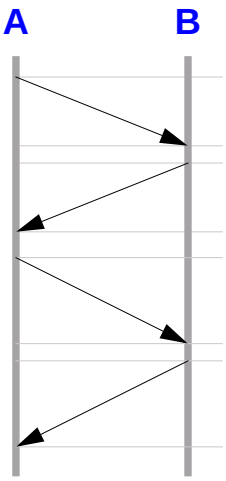
- C = Checksumme zur Überprüfung der richtigen Entschlüsselung
- A kann mehrere Verbindungen mit seinem Ticket öffnen

4 Beispiel: Kerberos (4)

- Unterscheidung zwischen Benutzer *(User)* und Benutzerprogramm *(Client)*
  - Wie kann ein Benutzerprogramm seinen Benutzer identifizieren?
  - Geheimer Schlüssel vom Benutzer (K_X) hängt von einem Paßwort ab
  - mittels einer Einwegfunktion wird aus dem Paßwort der Schlüssel K_X erzeugt
  - Benutzerprogramm braucht also das Paßwort zur Verbindungsaufnahme
- Beispiel: *kinit, klogin*
  - Anmeldung beim Authentisierungsdienst mit *kinit* und Paßworteingabe
  - Ticket wird im Benutzerkatalog gespeichert
  - *klogin* erlaubt das Einloggen auf einem entfernten Rechner mit Datenverschlüsselung und ohne Paßwort

5 Austausch öffentlicher Schlüssel

- A und B tauschen ihre öffentlichen Schlüssel aus



A schickt an B seinen öffentlichen Schlüssel: K_A

B schickt an A seinen öffentlichen Schlüssel: K_B

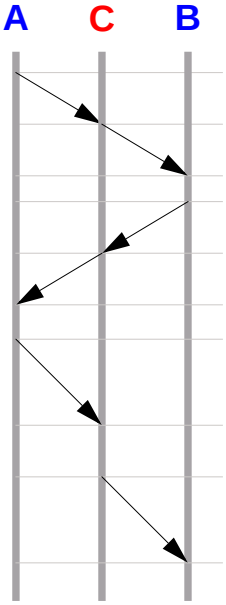
A verschickt signierte und verschlüsselte Nachricht:
$E(K_B, E(K_A^{-1}, T_1))$

B verschickt signierte und verschlüsselte Nachricht:
$E(K_A, E(K_B^{-1}, T_2))$

- Problem
  - A und B können nicht sicher sein, daß der öffentliche Schlüssel wirklich vom jeweils anderen stammt

5 Austausch öffentlicher Schlüssel (2)

- Aktiver Mithörer C fängt Datenverbindungen ab *(Man in the middle attack)*



A versucht an B seinen öffentlichen Schlüssel zu schicken: K_A

C fängt Nachricht ab und schickt an B seinen eigenen öffentlichen Schlüssel: K_C

B versucht an A seinen öffentlichen Schlüssel zu schicken: K_B

C fängt Nachricht ab und schickt an A seinen eigenen öffentlichen Schlüssel: K_C

A versucht an A eine Nachricht zu schicken:
$C_1 = E(K_C, E(K_A^{-1}, T_1))$

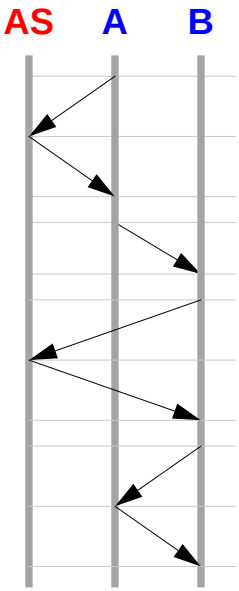
C dekodiert die Nachricht: $T_1 = D(K_C^{-1}, D(K_A, C_1))$

C schickt die Nachricht weiter: $E(K_B, E(K_C^{-1}, T_1))$

B glaubt eine Nachricht von B zu empfangen

5 Austausch öffentlicher Schlüssel (3)

- Einsatz eines Authentisierungsdienstes



A kontaktet AS und möchte Verbindung mit B: a, b

AS verschickt den Schlüssel von B: $E(K_{AS}^{-1}, <K_B, b>)$

A schickt initiale Nachricht an B: $E(K_B, <a, i_a>)$

B kontaktet AS und möchte den Schlüssel von A: b, a

AS verschickt den Schlüssel von A: $E(K_{AS}^{-1}, <K_A, a>)$

B antwortet an A: $E(K_A, <i_a, i_b>)$

A bestätigt: $E(K_B, <i_b>)$

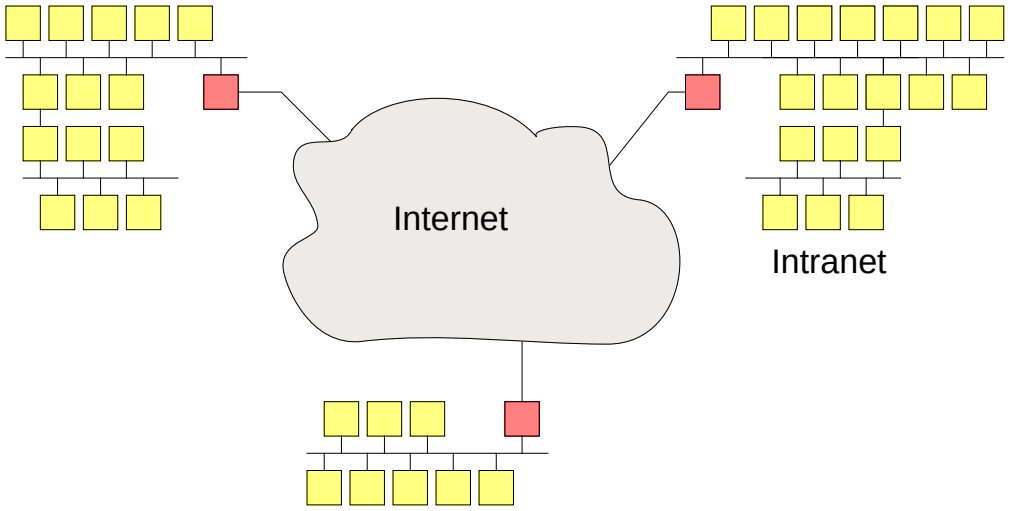
- Replay-Probleme
  - Hinzunahme von Zeitstempel und Lebendauer

I.8 Firewall

- Trennung von vertrauenswürdigen und nicht vertrauenswürdigen Netzwerksegmenten durch spezielle Hardware *(Firewall)*
  - Beispiel: Trennen des firmeninternen Netzwerks (Intranet) vom allgemeinen Internet
- Funktionalität
  - Einschränkung von Diensten
    - von innen nach außen, z.B. nur Standarddienste
    - von außen nach innen, z.B. kein Telnet, nur WWW
  - Paketfilter
    - Filtern „defekter“ Pakete, z.B. SYN-Pakete
  - Inhaltsfilter
    - Filtern von Pornomaterial aus dem WWW oder News
  - Authentisieren von Benutzern vor der Nutzung von Diensten

I.8 Firewall (2)

- Virtual private network
  - ◆ Verbinden von Intranet-Inseln durch spezielle Tunnels zwischen Firewalls
  - ◆ getunnelter Datenverkehr wird verschlüsselt
  - ◆ Benutzer sieht ein „großes Intranet“ (nur virtuell vorhanden)



I.9 Richtlinien für den Benutzer

1 Paßwörter

- Wahl eines Paßworts
  - ◆ hinreichend komplexe Paßwörter wählen
  - ◆ Schutz vor Wörterbuchangriffen
  - ◆ verschiedene Paßwörter für verschiedene Aufgaben
    (z.B. PPP Paßwort ungleich Benutzerpaßwort)
- Aufbewahrung
  - ◆ möglichst nirgends aufschreiben
  - ◆ nicht weitergeben
  - ◆ kein Abspeichern auf einem Windows-Rechner (Option immer wegklicken)

1 Paßwörter

- Eingabe
  - ◆ niemals über eine unsichere Rechnerverbindung eingeben
    - **`ftp, telnet, rlogin`** Dienste vermeiden
    - nur sichere Dienste verwenden: **`ssh, slogin`**
    - Datenweg beachten, über den das Paßwort läuft: ein unsicheres Netzwerk ist bereits genug
- Änderung
  - ◆ Paßwörter regelmäßig wechseln
  - ◆ alte Paßwörter nicht wiederverwenden

2 Schlüsselhandhabung

- Einsatz von PGP oder S/MIME
  - ◆ Zugang zu den privaten Schlüsseln für andere verhindern
  - ◆ Dateirechte auf der Schlüsseldatei prüfen
  - ◆ privater Schlüssel nur auf Diskette
  - ◆ Passphrase wie ein Paßwort behandeln
  - ◆ privaten Schlüssel nie über unsichere Netze transportieren

3 E-Mail

- Authentisierung
  - Bei elektronischer Post ist der Absender nicht authentisierbar
  - Digitale Unterschriften einsetzen (z.B. mit PGP oder S/MIME)
- Abhören
  - Elektronische Post durchläuft viele Zwischenstationen und kann dort jeweils gelesen und verfälscht werden
  - Verschlüsselung einsetzen (z.B. mit PGP oder S/MIME)

4 Programmierung

- S-Bit Programme vermeiden
  - Oft kann das S-Bit durch geschickte Vergabe von Benutzergruppen an Dateien vermieden werden
- Verwendung zusätzlicher Rechte (z.B. durch S-Bit) nur in Abschnitten
  - Trusted Computing Base (TCB)

```
seteuid(getuid());/* am Programmanfang Rechte wegnehmen */
...
seteuid(0);       /* setzt root Rechte */
fd = open("/etc/passwd", O_RDWR);
seteuid(getuid());/* nimmt root Rechte wieder weg */
...
```

- Sorgfältige Programmierung
  - Funktionen wie **`strcpy, strcat, gets, sprintf, scanf, sscanf, system, popen`** vermeiden oder durch **`strncpy, fgets, snprintf`** ersetzen

5 World Wide Web

- Cookies
  - Akzeptieren von Cookies erlaubt einer Web site die angesprochenen Seiten genau einem Benutzer zuzuordnen
  - funktioniert über Sessions hinweg
- JavaScript
  - schwere Sicherheitslücken erlauben es, alle für den Benutzer lesbare Dateien an einen dritten weiterzugeben
  - Ausschalten!
- Abhören
  - WWW Verbindungen können abgehört werden
  - keine privaten Daten, wie z.B. Kreditkartennummern übertragen
  - HTTPS mit SSL-Verschlüsselung benutzen; https URLs

J Mach

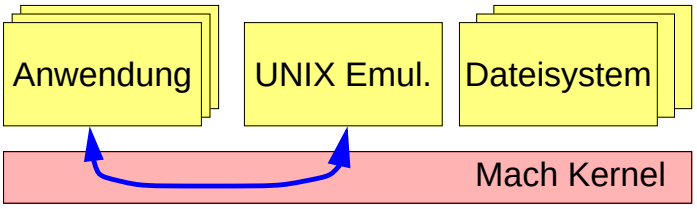
- Fallbeispiel: *Mach* Betriebssystem
  - Mikrokern
  - Unterstützung für Multiprozessoren
  - neues, hardwareunabhängiges Konzept der virtuellen Speicherverwaltung
  - *Capability*-basierte Interprozeßkommunikation – transparent erweiterbar für Netzwerk-Kommunikation
  - Modularer Aufbau, einfache Erweiterbarkeit
    (nur die wichtigsten Funktionen sind im Mach-Kern realisiert)
  - Betriebssystemumgebung wird durch eine Reihe von *Servern* realisiert
- Geschichte
  - Entwicklung 1985–1994 an der Carnegie Mellon Univ. (CMU)
  - Basis für OSF/1 (Open Software Foundation), NeXT-OS (Next/Apple)
  - Bedeutung heute eher gering

J.1 Motivation

- Entwicklung des UNIX-Systemkerns
  - heute nicht mehr so einfach aufgebaut und leicht modifizierbar wie früher
  - früher auf PDP11 mit 32k HSP lauffähig, heute 1,5–2 MB Speicher (inkl. Puffer etc.) notwendig
  - viele neue Funktionen werden kritiklos in den Kern eingebaut, weil dort die notwendigen Information leicht zugänglich sind
  - Aufblähung des Kerns, Abstraktionen und Strukturierung werden mehr und mehr zerstört
  - für Einsatz von UNIX auf Multiprozessorsystemen sind aufwendige Modifikationen notwendig
    - Kern-Monitor muß in Teile zerlegt werden oder ganz aufgegeben werden
    - an Stellen, an denen die Monitor-Eigenschaft aufgegeben wird, ist aufwendige Koordinierung erforderlich

J.1 Motivation (2)

- Architektur von Mach
  - Mikrokern
  - Betriebssystemumgebung wird durch eine Reihe von *Servern* realisiert, die über die Basis-Mechanismen des Mach-Kerns angesprochen werden können



  - Beispiele:
    - Dateisystem
    - Netzwerk-Kommunikation (z.B. TCP/IP)
    - Scheduling
    - UNIX-Emulation

J.2 Abstraktionen des Mach-Kerns

- *Task:* Ablaufumgebung für *Threads*
  - virtueller Adreßraum
  - gesicherter Zugriff zu Systemressourcen (Prozessoren, *Port-Capabilities*, Speicher)
- *Thread:* Aktivitätsträger innerhalb einer *Task*
  - beschreibbar durch
    - einen unabhängigen Programmzähler
    - eigenen Stack-Bereich innerhalb des virt. Adreßraums der *Task*
  - alle *Threads* einer *Task* haben gemeinsamen Zugriff zu allen *Task*-Ressourcen
- *Port:* Kommunikationskanal mit Warteschlange für Nachrichten
  - wird vom MACH-Kern verwaltet

J.2 Abstraktionen des Mach-Kerns (2)

- *Message:* Menge von Datenobjekten für die Kommunikation zwischen *Threads*
  - Messages können typisiert werden
  - können max. den gesamten virt. Adreßraum umfassen
  - können Zeiger und *Capabilities* für *Ports* enthalten
- Operationen
  - Operationen auf einem Objekt (ausgenommen *Messages*) werden durch Versenden von *Messages* an *Ports* ausgeführt, die dieses Objekt repräsentieren.
  - Beispiel:
    - Beim Generieren einer *Task* erhält man die Zugriffsrechte auf einen *Port* dieser *Task*.
    - Durch *Messages* an diesen *Port* kann die *Task* beeinflußt werden.

J.3 Tasks und Threads

1 Task

- Betriebsumgebung (System-Ressourcen) für Aktivitätsträger
  - ◆ virtueller Adreßraum
  - ◆ Zugriffsrechte
  - ◆ Betriebsmittel-Informationen
  - ◆ kein Programmablauf und keine Register

2 Thread

- Aktivitätsträger und seine Ablaufumgebung
  - ◆ Registersatz
  - ◆ Stack
  - ◆ Programmzähler
- Innerhalb einer *Task* können beliebig viele *Threads* existieren
  - ◆ In einer echten Multiprozessorumgebung können verschiedene *Threads* einer *Task* parallel auf mehreren Prozessoren ablaufen
  - ◆ Ein herkömmlicher UNIX-Prozeß entspricht einer MACH-*Task* mit einem *Thread*

3 Interprocess Communication – IPC

- nachrichtenorientierte Kommunikation
- Port
  - ◆ geschütztes Objekt des MACH-Kerns
  - ◆ kann Nachrichten *(Messages)* von einem Sender entgegennehmen und an einen Empfänger ausliefern
  - ◆ zu einem *Port* können beliebig viele Sender, aber nur ein Empfänger existieren
  - ◆ Zugriff auf einen *Port* erhält man durch Empfang einer Mitteilung, die ein *Port*-Zugriffsrecht *(Capability)* enthält

3 Interprocess Communication – IPC (2)

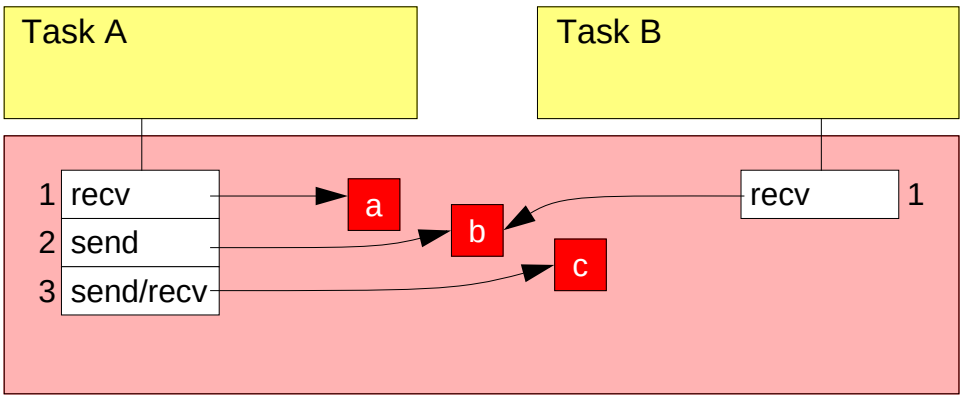
- Message
  - ◆ Ansammlung von Datenobjekten
  - ◆ unstrukturiert oder strukturiert (durch einen Typ näher gekennzeichnet), z.B.:
    - Daten mit Typinformation
    - Zeiger auf Adreßraumbereiche
    - *Port*-Zugriffsrechte *(Capabilities)*
- IPC ist in MACH der allgemeine Mechanismus zur Ausführung von Operationen auf Objekten
  - ◆ Objekten (z. B. *Tasks* und *Threads*) sind Ports zugeordnet
  - ◆ Operationen auf Objekten werden durch Nachrichten an diese Ports ausgeführt
  - ◆ zuverlässige Nachrichtenübertragung

3 Interprocess Communication – IPC (3)

- Beispiel
  - Beim Erzeugen eines Threads erhält man Senderechte auf die *Ports* dieses *Threads*. Durch Mitteilungen an diese *Ports* kann der *Thread* manipuliert werden (suspendieren, deblockieren, vernichten, ...)
- Ablauf einer typischen Interaktion
  - Ein Thread von Task A schickt eine Nachricht an einen Port
  - Task B ist der Empfänger für diesen Port

3 Interprocess Communication – IPC (4)

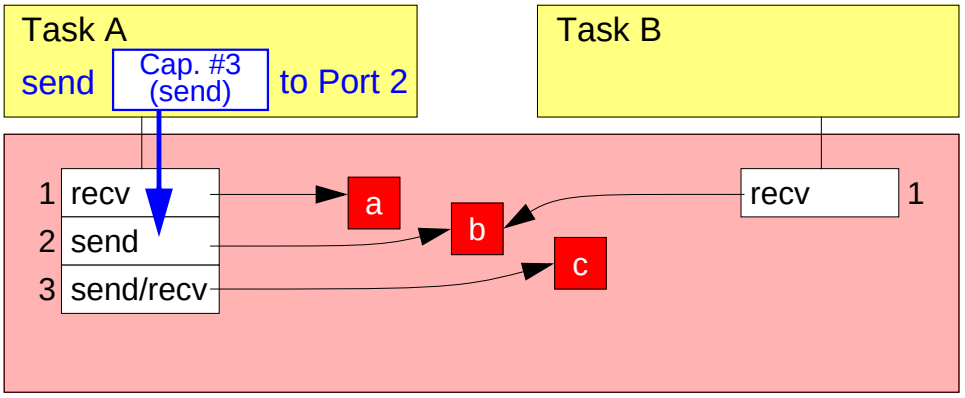
- Capabilities werden ähnlich wie Dateideskriptoren verwaltet



  - Jeder Task hat einen privaten Deskriptor für eine Portcapability
  - Deskriptor ist mit Rechten verbunden:
    Senderecht (send), Empfangsrecht (recv)

3 Interprocess Communication – IPC (5)

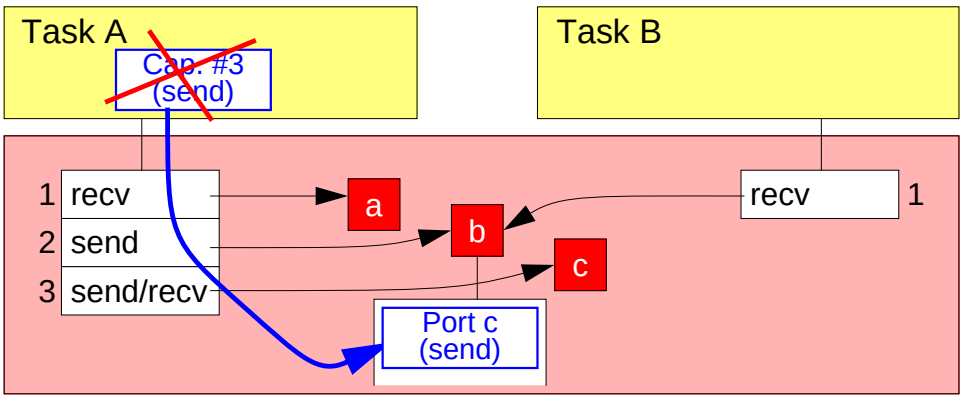
- Task A erzeugt eine Nachricht und sendet sie



◆ Adressat ist Port mit Deskriptor 2, Port b

3 Interprocess Communication – IPC (6)

- Einstellen der Nachricht in der Port-Warteschlange



◆ Nachricht wird in die Warteschlange des Port b eingefügt

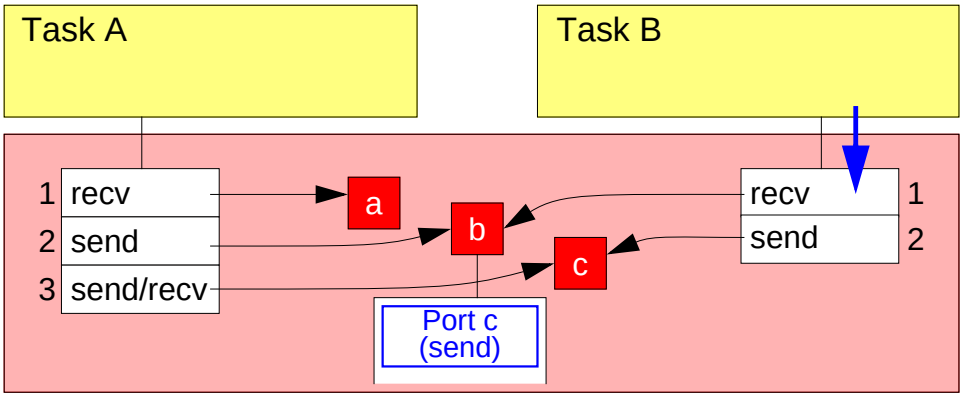
◆ übermittelte Capability wird in Portadresse umgesetzt

◆ Nachricht steht jetzt zum Empfang bereit

◆ Task A kann den Nachrichtenspeicher löschen oder wiederverwenden

3 Interprocess Communication – IPC (7)

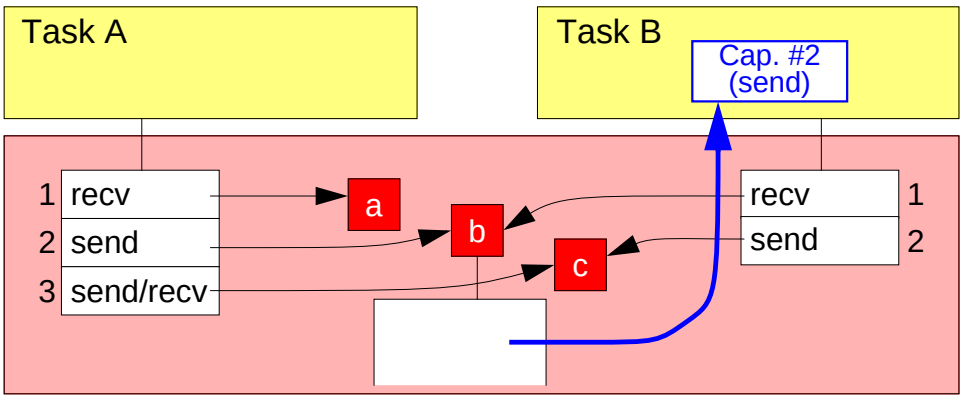
- Task B möchte eine Nachricht empfangen



- Task B benötigt neuen Deskriptor für die Capability in der Nachricht
- Deskriptor 2 mit dem entsprechenden Recht wird erzeugt

3 Interprocess Communication – IPC (8)

- Task B empfängt Nachricht



- Task B erhält die Nachricht
- Portadresse wird in Deskriptor übersetzt
- Nachricht wird aus der Warteschlange entfernt
- Task B hat nun Zugriff auf den Port c

4 Übertragen einer „großen“ Nachricht

- Ablauf
  - Task A sendet Message an *Port* P1
  - der Datenbereich der *Message* wird temporär in den Adreßraum des Mach-Kerns abgebildet (= Senden an *Port*)
  - im Adreßraum der Task A wird der Datenbereich als *”copy on write”* markiert (d. h. für Seiten, auf die nachfolgend schreibend durch A zugegriffen wird, wird eine Kopie erzeugt und diese wird in den virtuellen Adreßraum von A übernommen)
  - Task B empfängt die Mitteilung, indem der Datenbereich in den virt. Adreßraum von B abgebildet wird
  - Aus dem Adreßraum des Mach-Kerns wird der Datenbereich wieder entfernt
  - Auch bei *Task* B ist der Datenbereich als *”copy on write”* markiert

J.4 Virtuelle Speicherverwaltung (VM)

- Eigenschaften
  - portabel
  - hardwareunabhängig
  - unterstützt Multiprozessoren: parallel ablauffähig implementiert
  - unabhängige *Pager*

1 Motivation

- heute viele unterschiedliche MMUs verbreitet
- weitgehend äquivalente Funktionalität
  - unterschiedliche virtuelle Adreßräume, beschrieben durch Datenstrukturen (z. B. *Page table*)
  - pro Seite:
    - Zeiger auf physikalische Adresse der Seite
    - *Valid bit* zeigt an, ob die physikalische Seite gültig ist
    - *Protection bits* (z. B. *read/write, user/kernel*)
    - *Dirty bit* (Seite ist modifiziert)
    - *Reference bit* (Seite wurde angesprochen)

1 Motivation (2)

- Zwei mögliche Vorgehensweisen zur Unterstützung von virtuellem Speicher
  - Mechanismen, die möglichst gut den Eigenschaften der verwendeten MMU entsprechen
    - System nicht portabel (z. B. VMS für VAX)
    - effizient
  - Beschränkung auf Eigenschaften, die einfach auf jeder Hardware realisierbar sind
    - System weniger leistungsfähig und ineffizient (z. B. UNIX in 4.3bsd)
    - aber portabler

2 Realisierung

- Ziel
  - ◆ möglichst flexible Unterstützung von virtuellem Speicher auf möglichst vielen Architekturen
- Vorgehensweise
  - ◆ Unterteilung der Implementierung
    - architekturabhängiger Teil
    - architekturunabhängiger Teil
    - externe Pager

2 Realisierung (2)

- Architekturabhängiger Teil
  - ◆ Verwaltung der von der Hardware benötigten Adreßabbildungstabellen *(Physical address maps, pmaps)*
    - entsprechen den MMU-Tabellen (z. B. *VAX page table* oder eine Menge von Segmentregistern beim IBM PC/RT)
  - ◆ Abbildung von Mach-Seiten in eine oder mehrere Kacheln
  - ◆ konstruiert nur einen Teil der Adreßabbildung
  - ◆ Zugriff auf weitere Adressen kann bei *Page fault* ermöglicht werden
  - ◆ kleine Tabellen trotz großem (dünn besetztem) Adreßraum

2 Realisierung (3)

- Architekturunabhängiger Teil
  - ◆ realisiert Benutzerschnittstelle und Funktionsumfang
  - ◆ Verwaltung der architekturunabhängigen Datenstrukturen
  - ◆ alle notwendigen Abbildungsinformationen für jede Seite des Systems können aus den architekturunabhängigen Datenstrukturen gewonnen werden
- externe Pager
  - ◆ Seitenein- und -auslagerung wird durch die virtuelle Speicherverwaltung (VM) von sogenannten *Pagern* über Messages angefordert
    - entscheiden nicht über Ein- und Auslagerungsstrategie
    - entscheiden wohin ausgelagerte Seiten transportiert und woher eingelagerte Seiten bezogen werden (z.B. Platte, Netzwerk etc.)

2 Realisierung (4)

- Lazy evaluation
  - ◆ Speicher (Haupt- und Hintergrundspeicher!) wird erst belegt, wenn eine Seite benutzt wird
  - ◆ *Map-on-Reference:* Page tables werden nur für aktuell benötigten Speicher bereitgestellt
  - ◆ *Copy-on-write:* Seiten werden solange gemeinsam benutzt wie möglich; bei modifizierendem Zugriff werden Nutzer voneinander getrennt

3 Datenstrukturen

- Memory Object
  - repräsentiert einen Speicherbereich, der ganz oder teilweise in einen Adreßraum eingeblendet werden kann
  - Bindeglied zwischen virtuellem Adreßraum und Hintergrundspeicher
  - das *Memory object* kennt oder implementiert einen Pager, der weiß woher die Seiten des Speicherbereichs beschafft werden können und was im Falle eines *Pagout* zu tun ist
  - Pager (intern oder extern) realisiert Ein-/Auslagerung von Seiten
  - *Memory object* hält Liste von Seiten, die momentan im Hauptspeicher vorhanden sind *(cached pages)*

3 Datenstrukturen (2)

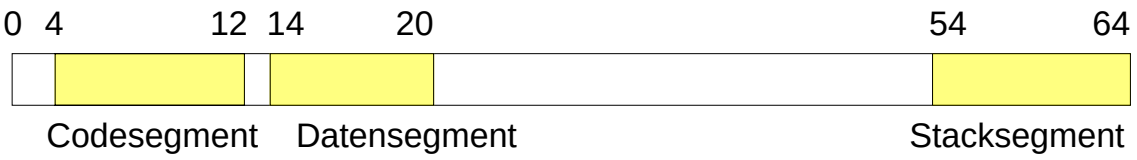
- Address Map
  - repräsentiert einen Adreßraum
  - als verkettete Liste von *Address map entries* realisiert
  - jeder Eintrag bildet einen (virtuellen) Speicherbereich mit gleichen Eigenschaften (Schutz, Vererbung) in einen Abschnitt eines *Memory objects* ab
  - *Address map entry* enthält:
    - Vorwärts- und Rückwärts-Verzeigerung
    - Start- und Endadresse des Speicherbereichs im virtuellen Adreßraum
    - aktuelle und maximale Zugriffsrechte
    - Regel für Vererbung des Speicherbereichs an neue Task
    - Zeiger auf *Memory object*
    - Offset dieses Speicherbereichs im entsprechenden *Memory object*

3 Datenstrukturen (3)

- Sharing Maps
  - zur Realisierung von *Shared memory*
  - Zwischenstufe zwischen *Address map* und *Memory object*
  - Änderungen im *Address map*-Eintrag eines Objekts müssen nicht in den *Address maps* mehrerer Tasks nachgetragen werden
- Shadow Object
  - wird wie ein *Memory object* benutzt
  - zur Realisierung von *Copy-on-write*
  - nimmt Seiten auf, die kopiert werden mußten
  - verweist für unveränderte Seiten auf das Originalobjekt

3 Datenstrukturen (4)

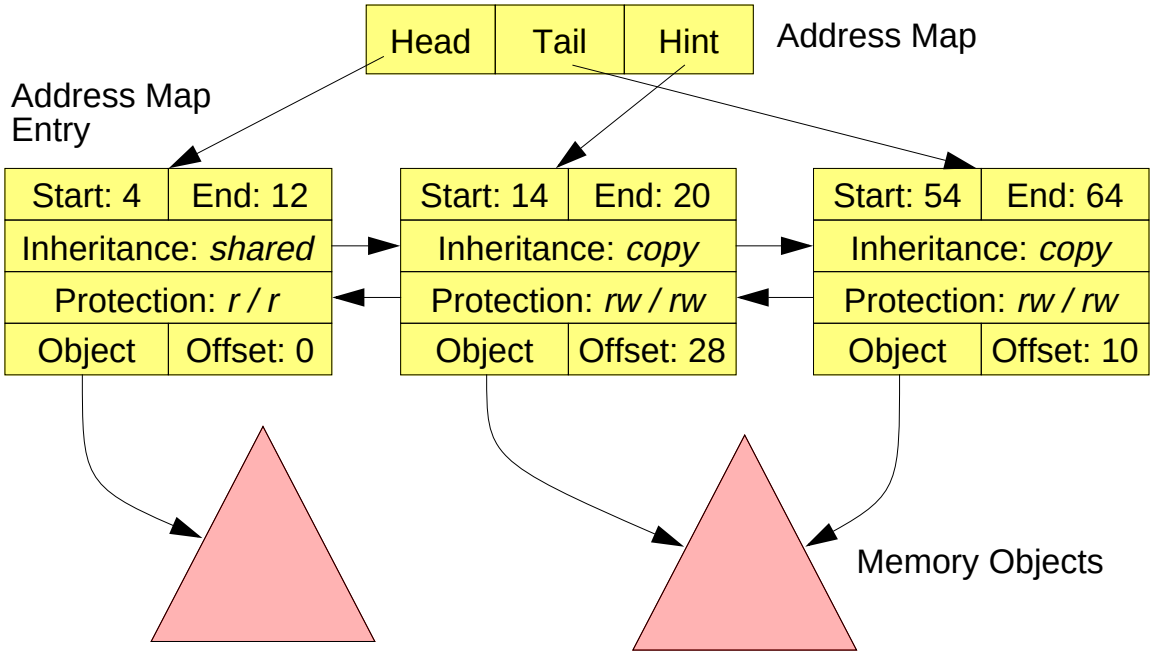
- Resident Page Table
  - Verwaltung des Hauptspeicher-Caches für *Memory object*-Seiten
  - Verkettung in Listen
- Beispiel eines Adreßraums (z.B. einer Task)



  - Abbildung des Codesegments in eine Programmdatei
  - Abbildung des Daten- und Stacksegments in einen *Swap space*

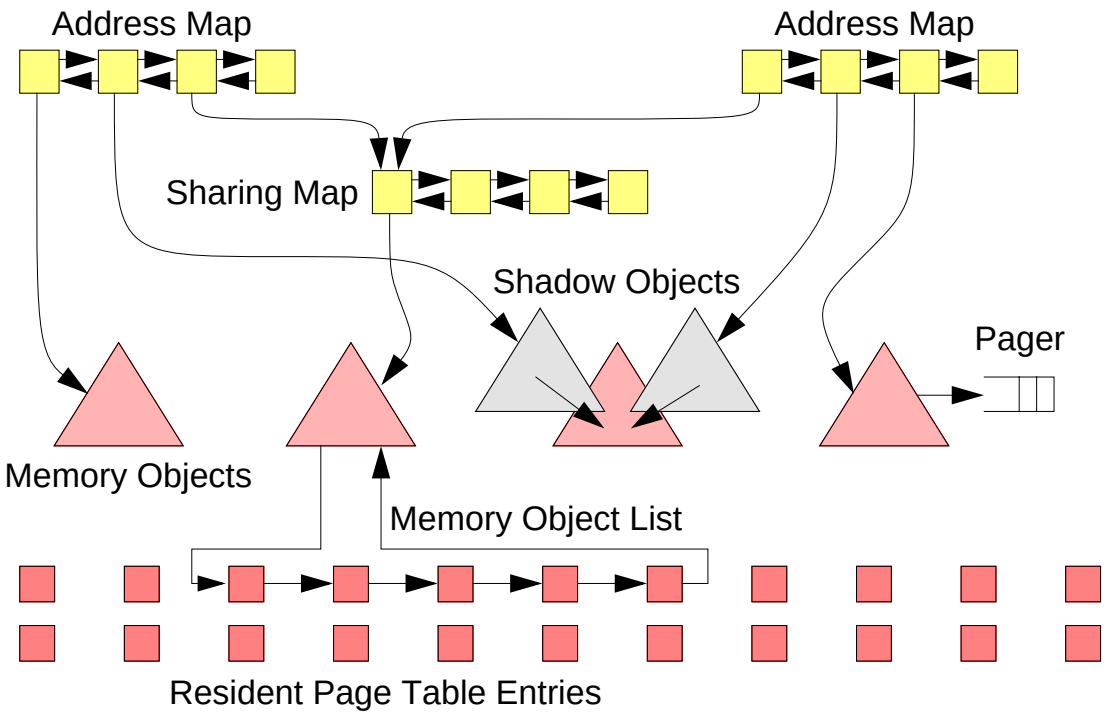
3 Datenstrukturen (5)

- Datenstrukturen der Adreßraumabbildung: *Address map*



3 Datenstrukturen (6)

- Beispielhafter Gesamtüberblick



4 Benutzerschnittstelle

- MACH-VM stellt folgende Funktionen zur Verfügung:
  - ◆ virtuelle Speicherbereiche belegen und freigeben
  - ◆ Schutzparameter für einzelne virtuelle Speicherbereiche individuell einstellen (innerhalb vorgegebener Schranken: *current / maximum protection*)
  - ◆ Vererbung an neue Task einzeln einstellbar
    - keine Vererbung
    - Copy-on-write (default)
    - Shared
  - ◆ Zugriffsmöglichkeit auf Speicherabbildung anderer Tasks (Debugger!)
  - ◆ Status- und Statistikabfragen

4 Benutzerschnittstelle (2)

- Anwendungsbeispiel
  - ◆ *Memory mapped files* für Benutzer und Kern
  - ◆ Realisierung: Memory object mit entsprechendem Pager
  - ◆ Kern-Anwendungen:
    - Programme laden
    - *Shared libraries*
    - Dateisystem
  - ◆ Benutzeranwendungen:
    - neue *Stdio* mit Dateizugriff über *Mapped file*
    - Dateizugriff über Pointer

4 Benutzerschnittstelle (3)

- Erfahrungen mit verschiedenen Hardware-Architekturen
  - Virtuelle Speicherverwaltung von Mach ist
    - portabel
    - stellt wenig Anforderungen an die Hardware-Plattform
    - und wurde auf einer Reihe von Architekturen implementiert
  - gute Basis für Untersuchungen von VM-Problemen bei verschiedenen Architekturen

5 VM auf Multiprozessoren

- Problem: *Translation Lookaside Buffer (TLB)* -Inkonsistenz
  - TLB = Cache für die Adreßdaten der SKT, um eine schnelle Adreßübersetzung ohne zusätzliche Hauptspeicher-Zugriffe durch die MMU zu ermöglichen
  - bei *Shared memory*-Multiprozessoren werden normalerweise Strategien implementiert, um die Caches der Prozessoren konsistent zu halten
  - analog ist es notwendig die TLBs der Prozessoren konsistent zu halten, wenn Threads einer Task parallel auf verschiedenen Prozessoren ablaufen
    - wenn sich die architekturabhängigen VM-Datenstrukturen einer Task ändern, müssen die TLBs aller Prozessoren aktualisiert werden, auf denen Threads der Task ablaufen
    - in anderen Betriebssystemen ist dies meist kein Problem, weil in einem Adreßraum nicht mehrere Aktivitätsträger existieren
    - die *Lazy evaluation*-Techniken in Mach führen dazu, daß sehr häufig die pmap-Strukturen angepaßt werden müssen

5 VM auf Multiprozessoren (2)

- viele heutige Multiprozessorarchitekturen ignorieren dieses Problem!

■ Intuitive Lösungsidee

- nach einer Änderung der VM-Abbildung werden durch einen Befehl die TLBs der anderen Prozessoren invalidiert

▲ Problem

- die meisten MPS-Systeme erlauben es einem Prozessor nur, seinen eigenen TLB, aber nicht die anderer Prozessoren zu invalidieren

■ weitergehende Lösungsideen

- Hardware-Unterstützung implementieren oder
- durch Software-Interrupts die anderen Prozessoren zum Invalidieren ihrer TLBs bewegen

5 VM auf Multiprozessoren (3)

▲ Problem

- systeminitierte Aktualisierung der pmap-Strukturen überschneidet sich mit der prozessorinitierten Aktualisierung (z.B. Setzen des *Reference bit*)
- inkonsistente Menge von TLB-Einträgen in der MMU möglich
- Überschreiben der pmap-Strukturen durch die MMU möglich

■ Mögliche Lösungen

- Prozessoren werden durch Interrupt informiert und führen ein TLB-Konsistenz-Protokoll durch
- die Nutzung geänderter Abbildungen wird verzögert, bis alle TLBs *geflushed* wurden
- Inkonsistenzen werden in unkritischen Fällen toleriert (z. B. wenn die Zugriffsrechte erweitert werden)
- Komplexere Algorithmen: z. B. *Shootdown* Algorithmus

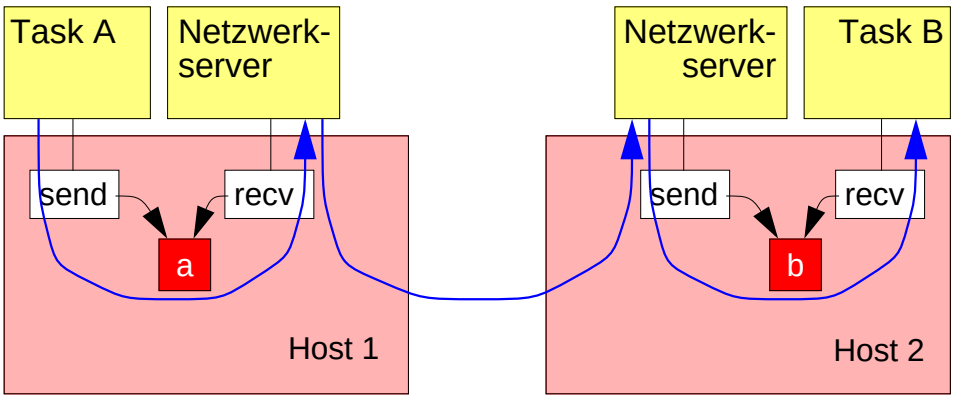
J.5 Mach in verteilten Systemen

1 Netzwerk-Kommunikation

- Mach IPC nur für lokale Kommunikation konzipiert
  - Nachrichten können durch *Network-Server-Tasks* (im User-level!) an andere Knoten weitergeleitet werden
  - Capability-Funktionalität wird über Netzwerk abgebildet
  - Netzwerktransfer wird geschützt
    - DES-Verschlüsselung der Daten
    - Sequenznummern und Checksummen für Pakete
  - Authentisierung anderer Netzwerk-Server für den Capability-Transfer (*Secret Identifier*)

1 Netzwerkkommunikation (2)

- Schaubild einer verteilungstransparenten Taskkommunikation



  - Netzwerkserver fungieren als Vermittler
  - sie stellen jeweils lokale Ports bereit und tunneln Kommunikation durch die Ports über ein Netzwerk

2 Distributed Shared Memory

- der *Shared memory*-Bereich wird durch ein *Memory Object* repräsentiert
  - jede beteiligte Task bildet den Speicherbereich in ihren Adreßraum ab
  - dem *Memory Object* wird ein *External pager* zugeordnet
  - durch die *Lazy evaluation*-Technik werden die Seiten erst auf Anforderung in den Hauptspeicher (Seiten-Cache) geholt
  - im Fall von *Page faults* werden die Seiten von dem *External pager* angefordert
    - diese Anforderungen erfolgen über die regulären Mach Kommunikationsmechanismen, sind also auch auf Netzwerkkommunikation erweiterbar

2 Distributed Shared Memory (2)

- als *External pager* wird ein sog. ***Shared memory server*** benutzt, der Buch über das Caching der Seiten führt
- *read-only* Seiten können problemlos auf mehreren Knoten gecached werden
- bei Schreibzugriff wird ein *Page fault* erzeugt; der *Shared memory server* wird informiert, er entzieht den anderen Knoten den Zugriff auf die Seite und erlaubt dann den Schreibzugriff